

(19) 日本国特許庁(JP)

(12) 公表特許公報(A)

(11) 特許出願公表番号

特表2005-505037

(P2005-505037A)

(43) 公表日 平成17年2月17日(2005.2.17)

(51) Int. Cl. <sup>7</sup>	F I	テーマコード (参考)
G06F 13/10	G06F 13/10 340A	5B014
G06F 3/06	G06F 3/06 301A	5B065
G06F 12/00	G06F 3/06 301N	5B082
H04L 29/06	G06F 12/00 545A	5K034
	H04L 13/00 305B	

審査請求 未請求 予備審査請求 有 (全 171 頁)

(21) 出願番号 特願2003-531344 (P2003-531344)  
 (86) (22) 出願日 平成14年9月27日 (2002. 9. 27)  
 (85) 翻訳文提出日 平成16年3月29日 (2004. 3. 29)  
 (86) 国際出願番号 PCT/US2002/030912  
 (87) 国際公開番号 W02003/027877  
 (87) 国際公開日 平成15年4月3日 (2003. 4. 3)  
 (31) 優先権主張番号 60/325, 704  
 (32) 優先日 平成13年9月28日 (2001. 9. 28)  
 (33) 優先権主張国 米国 (US)  
 (31) 優先権主張番号 10/051, 415  
 (32) 優先日 平成14年1月18日 (2002. 1. 18)  
 (33) 優先権主張国 米国 (US)

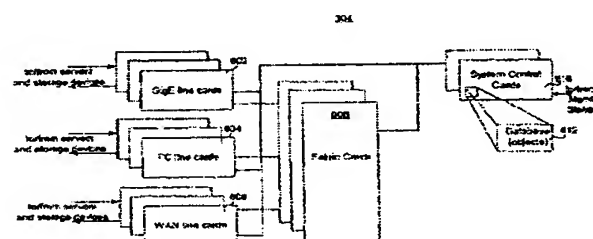
(71) 出願人 504121830  
 マランティ ネットワークス インコーポ  
 レイテッド  
 アメリカ合衆国 カリフォルニア州 95  
 134-2127 サン ホセ ザンカー  
 ロード 3061 ビー  
 (74) 代理人 100082005  
 弁理士 熊倉 禎男  
 (74) 代理人 100067013  
 弁理士 大塚 文昭  
 (74) 代理人 100074228  
 弁理士 今城 俊夫  
 (74) 代理人 100086771  
 弁理士 西島 孝喜

最終頁に続く

(54) 【発明の名称】 記憶システムにおけるプロトコル変換

## (57) 【要約】

本発明の実施形態による記憶装置スイッチは、SANの作成を可能にし、分散が容易で、中央管理が可能な拡張性の高いスイッチである。更に、この記憶装置スイッチは、世界的規模のインフラストラクチャの分散を可能にするので、記憶装置等のSANを実質的に世界のどこにでも配置することができる。更に、この記憶装置スイッチは、1 SCSI又はファイバ・チャネルの両方を含むもの等の多重プロトコルSANを可能にし、データパケットを「ワイヤ速度」で処理する。更なるワイヤ速度処理を可能にするために、本発明によるスイッチは、ラインカードの各々に分散された「知能」を有し、ラインカードによって、パケットをデータ及び制御パケットに分類し、仮想化機能を実行し、プロトコル変換機能を実行する。更に、本発明によるスイッチは、ミラーリング、スナップショット、複製等のサーバレス記憶サービスを実行する。



**【特許請求の範囲】****【請求項1】**

データを記憶してアクセスするためのシステムに使用方法であって、

(a) 第1のプロトコルに従ってフォーマットされたパケットを、前記第1のプロトコルに従って作動する第1の装置から受信するステップと、

(b) 前記パケットを第2のプロトコルに従ってフォーマットされたパケットに変換するステップと、

(c) 前記パケットを前記第2のプロトコルに従って作動する第2の装置に送信するステップと、

を含み、前記ステップ(a)から(c)は、前記パケットのバッファリングを行うことなく実行されることを特徴とする方法。

**【請求項2】**

前記ステップ(a)から(c)は、ワイヤ速度で実行されることを特徴とする請求項1に記載の方法。

**【請求項3】**

前記変換ステップは、前記第1のプロトコルに従ってフォーマットされた前記パケットのフィールドの少なくとも一部を、前記第2のプロトコルに従ってフォーマットされた前記パケットにマッピングするステップを含むことを特徴とする請求項1に記載の方法。

**【請求項4】**

前記変換ステップは、前記第1のプロトコルに従ってフォーマットされた前記パケットからマッピングされていない、前記第2のプロトコルに従ってフォーマットされた前記パケットの任意のフィールドに情報を追加するステップを更に含むことを特徴とする請求項3に記載の方法。

**【請求項5】**

前記第1のプロトコルはiSCSIであり、前記第2のプロトコルはファイバ・チャンネルであることを特徴とする請求項1に記載の方法。

**【請求項6】**

前記第1のプロトコルは、TCPに組み込まれたiSCSIであることを特徴と

する請求項5に記載の方法。

【請求項7】

前記第1のプロトコルはファイバ・チャンネルであり、前記第2のプロトコルはiSCSIであることを特徴とする請求項1に記載の方法。

【請求項8】

前記第2のプロトコルは、TCPに組み込まれたiSCSIであることを特徴とする請求項7に記載の方法。

【請求項9】

前記第1の装置はイニシエータであり、前記第2の装置はターゲットであることを特徴とする請求項1に記載の方法。

【請求項10】

前記第1の装置はターゲットであり、前記第2の装置はイニシエータであることを特徴とする請求項1に記載の方法。

【請求項11】

データを記憶してアクセスするためのシステムに使用方法であって、

(a) 第1のプロトコルに従ってフォーマットされたパケットを、前記第1のプロトコルに従って作動する第1の装置から受信するステップと、

(b) 前記パケットを第2のプロトコルに従ってフォーマットされたパケットに変換するステップと、

(c) 前記パケットを前記第2のプロトコルに従って作動する第2の装置に送信するステップと、

を含み、前記ステップ(a)から(c)は、ワイヤ速度で実行されることを特徴とする方法。

【請求項12】

データを記憶してアクセスするためのシステムに使用方法であって、

第1の認識プロトコルに従ってフォーマットされたパケットを受信するステップと、

前記パケットのバッファリングを行うことなく、前記パケットのフィールドを第2のプロトコルに従ってフォーマットされた新しいパケットにマッピングするス

テップと、  
を含むことを特徴とする方法。

**【請求項13】**

前記受信ステップ及び前記マッピングステップは、ワイヤ速度で実行されることを特徴とする請求項12に記載の方法。

**【請求項14】**

前記マッピングステップは、バッファリングを行うことなく、且つCPUリソースを使用することなく、前記パケットの前記フィールドを、第2のプロトコルに従ってフォーマットされた新しいパケットにマッピングするステップを含むことを特徴とする請求項12に記載の方法。

**【請求項15】**

前記第1のプロトコルはiSCSIであり、前記第2のプロトコルはファイバ・チャンネルであることを特徴とする請求項12に記載の方法。

**【請求項16】**

前記第1のプロトコルは、TCPに組み込まれたiSCSIであることを特徴とする請求項15に記載の方法。

**【請求項17】**

前記第1のプロトコルはファイバ・チャンネルであり、前記第2のプロトコルはiSCSIであることを特徴とする請求項12に記載の方法。

**【請求項18】**

前記第2のプロトコルは、TCPに組み込まれたiSCSIであることを特徴とする請求項17に記載の方法。

**【請求項19】**

前記パケットは、SCSIコマンド記述子ブロック(CDB)を含むことを特徴とする請求項12に記載の方法。

**【請求項20】**

データを記憶してアクセスするためのシステムに使用方法であって、

iSCSI又はファイバ・チャンネルプロトコルの1つに従ってフォーマットされたパケットを第1の装置から受信するステップと、



前記パッケージが前記 iSCSI プロトコルに従ってフォーマットされている場合、前記 iSCSI プロトコルに従ってフォーマットされた前記パッケージからのフィールドの少なくとも一部をファイバ・チャンネルパッケージにマッピングすることによって、前記パッケージを前記ファイバ・チャンネルパッケージに変換するステップと、

前記パッケージが前記ファイバ・チャンネルプロトコルに従ってフォーマットされている場合、前記ファイバ・チャンネルプロトコルに従ってフォーマットされた前記パッケージからのフィールドの一部を前記 iSCSI パッケージにマッピングすることによって、前記パッケージを前記 iSCSI パッケージに変換するステップと、

前記パッケージを変換された状態で第2の装置に送信するステップと、  
を含み、前記ステップの全ては、バッファリングを行うことなく実行されることを特徴とする方法。

【請求項21】

前記ステップの全ては、ワイヤ速度で実行されることを特徴とする請求項20に記載の方法。

【請求項22】

前記第1の装置はイニシエータであり、前記第2の装置はターゲットであることを特徴とする請求項20に記載の方法。

【請求項23】

前記第1の装置はターゲットであり、前記第2の装置はイニシエータであることを特徴とする請求項20に記載の方法。

【請求項24】

前記受信されたパッケージは iSCSI コマンド PDU であり、前記変換された状態のパッケージはファイバ・チャンネルコマンド IU であることを特徴とする請求項20に記載の方法。

【請求項25】

前記受信されたパッケージはファイバ・チャンネル XFR\_RDY IU であり、前記変換された状態のパッケージは iSCSI R2T PDU であることを特徴

とする請求項20に記載の方法。

【請求項26】

前記受信されたパケットはiSCSI書込みデータPDUであり、前記変換された状態のパケットはファイバ・チャンネルデータIUであることを特徴とする請求項20に記載の方法。

【請求項27】

前記受信されたパケットはファイバ・チャンネルデータIUであり、前記変換された状態のパケットはiSCSI読出しデータPDUであることを特徴とする請求項20に記載の方法。

【請求項28】

前記受信されたパケットはファイバ・チャンネル応答IUであり、前記変換された状態のパケットはiSCSI応答PDUであることを特徴とする請求項20に記載の方法。

【請求項29】

前記受信されたパケットはファイバ・チャンネルコマンドIUであり、前記変換された状態のパケットはiSCSIコマンドPDUであることを特徴とする請求項20に記載の方法。

【請求項30】

前記受信されたパケットはiSCSI R2T PDUであり、前記変換された状態のパケットはファイバ・チャンネルXFR\_\_RDY IUであることを特徴とする請求項20に記載の方法。

【請求項31】

前記受信されたパケットはファイバ・チャンネルデータIUであり、前記変換された状態のパケットはiSCSI書込みデータPDUであることを特徴とする請求項20に記載の方法。

【請求項32】

前記受信されたパケットはiSCSI読出しデータPDUであり、前記変換された状態のパケットはファイバ・チャンネルデータIUであることを特徴とする請求項20に記載の方法。

**【請求項33】**

前記受信されたパケットはiSCSI応答PDUであり、前記変換された状態のパケットはファイバ・チャンネル応答IUであることを特徴とする請求項20に記載の方法。

**【請求項34】**

データを記憶してアクセスするためのシステムに使用方法であって、  
仮想ターゲットアドレスで所定の仮想ターゲットに宛てられた、第1のプロトコルに従ってフォーマットされたパケットを入口側ラインカードにて受信するステップと、

前記入口側ラインカードが、フローIDを含む前記仮想ターゲットに関する情報を仮想ターゲット記述子から検索して、仮想ターゲット記述子識別子及び前記フローIDを前記パケットのローカルヘッダに付加するステップと、

前記入口側ラインカードが、前記フローIDに従って前記パケットを出口側ラインカードに送信するファブリックへ前記パケットを送信するステップと、

前記出口側ラインカードが、前記仮想ターゲット記述子識別子を使用して、前記仮想ターゲットに関連する物理的ターゲットが第2のプロトコルに従ってフォーマットされたパケットを必要とするか否かを含む、前記物理的ターゲットに関する情報を識別すると共に、前記物理的ターゲットに関する前記情報を使用して、仮想ターゲットブロックアドレスを物理的ターゲットブロックアドレスに変換し、必要であれば、前記パケットのフォーマットを前記第1のプロトコルから前記第2のプロトコルに変換するステップと、

前記出口側ラインカードが、前記物理的ターゲットブロックアドレスを使用して、前記パケットを前記物理的ターゲットに送信するステップと、  
を含むことを特徴とする方法。

**【請求項35】**

前記ステップの全ては、バッファリングを行うことなく実行されることを特徴とする請求項34に記載の方法。

**【請求項36】**

前記ステップの全ては、ワイヤ速度で実行されることを特徴とする請求項34に

記載の方法。

【請求項37】

ポートと、

前記ポートと通信を行い、バッファリングを行うことなくパケットを第1のプロトコルから第2のプロトコルに変換する変換器と、  
を備えることを特徴とするラインカード。

【請求項38】

前記変換器とは別個のCPUを更に含むことを特徴とする請求項37に記載のラインカード。

【請求項39】

前記変換は、前記第1のプロトコルに従ってフォーマットされた場合の前記パケットからのフィールドを、前記第2のプロトコルに従ってフォーマットされた場合の前記パケットのフィールドにマッピングするステップを含むことを特徴とする請求項37に記載のラインカード。

【請求項40】

前記第1のプロトコルはiSCSIであり、前記第2のプロトコルはファイバ・チャンネルであることを特徴とする請求項37に記載のラインカード。

【請求項41】

前記第1のプロトコルはファイバ・チャンネルであり、前記第1のプロトコルはiSCSIであることを特徴とする請求項37に記載のラインカード。

【請求項42】

ポートと、

前記ポートと通信を行い、バッファリングを行うことなくパケットを第1のプロトコルから第2のプロトコルに変換するための手段と、  
を備えることを特徴とするラインカード。

【請求項43】

第1のプロトコルに従って作動する第1の装置に結合されることができる第1のポートを有する第1のラインカードと、  
前記第1のラインカードと通信を行うファブリックと、

前記ファブリックと通信を行うと共に、第2のプロトコルに従って作動する第2の装置に結合されることができる第2のポートを有する第2のラインカードと、を備えるスイッチであって、

前記第2のラインカードは、プロトコル変換器を含み、前記プロトコル変換器は、前記第1のプロトコルに従ってフォーマットされたパケットを前記ファブリックから受信するために結合された入力部と、前記第2のポートに結合され、前記第2のプロトコルに従ってフォーマットされると共に前記第1のプロトコルに従ってフォーマットされた前記パケットに対応するパケットを生成する出力部とを有し、

前記第2のラインカードは、前記プロトコル変換器とは別個のCPUを更に含むことを特徴とするスイッチ。

【請求項44】

前記スイッチは、前記第1のポートにおいて前記第1のプロトコルに従ってフォーマットされたパケットを受信し、前記第2のポートにおいて前記第2のプロトコルに従ってフォーマットされた前記パケットをワイヤ速度で生成するように設計されることを特徴とする請求項43に記載のスイッチ。

【請求項45】

前記プロトコル変換器は、前記パケットのバッファリングを行うことなく作動することを特徴とする請求項43に記載のスイッチ。

【請求項46】

前記第1の装置はイニシエータであり、前記第2の装置はターゲットであることを特徴とする請求項43に記載のスイッチ。

【請求項47】

前記第1の装置はターゲットであり、前記第2の装置はイニシエータであることを特徴とする請求項43に記載のスイッチ。

【請求項48】

各々が少なくとも1つのポートを含む複数のラインカードと、第1のラインカード上の第1のポートにおいて、第1のプロトコルに従ってフォーマットされた特定の仮想ターゲット宛のパケットを受信し、第2のラインカー

ド上の第2のポートにおいて、第2のプロトコルに従ってフォーマットされた対応するパケットを前記仮想ターゲットに関連する物理的ターゲットにワイヤ速度で送信するための手段と、

を備えることを特徴とするスイッチ。

【請求項49】

プロセッサによって実行可能であり、データを記憶してアクセスするためのシステムに使用するスイッチ内の少なくとも1つの媒体上に記憶されたソフトウェア命令集合であって、

iSCSI又はファイバ・チャンネルプロトコルの1つに従ってフォーマットされたパケットを第1の装置から受信するための命令と、

前記パケットが前記iSCSIプロトコルに従ってフォーマットされる場合、前記iSCSIプロトコルに従ってフォーマットされた前記パケットからのフィールドの少なくとも一部をファイバ・チャンネルパケットにマッピングすることによって、前記パケットを前記ファイバ・チャンネルパケットに変換するための命令と、

前記パケットが前記ファイバ・チャンネルプロトコルに従ってフォーマットされる場合、前記ファイバ・チャンネルプロトコルに従ってフォーマットされた前記パケットからのフィールドの少なくとも一部をiSCSIパケットにマッピングすることによって、前記パケットを前記iSCSIパケットに変換するための命令と、

前記パケットを変換された状態で第2の装置に送信するための命令と、

を備え、前記命令の全ては、前記パケットのバッファリングを行うことを必要としないことを特徴とする命令。

**【発明の詳細な説明】****【技術分野】****【0001】**

本発明は、記憶領域ネットワーク（SAN）に関する。

**【背景技術】****【0002】**

集中的なデータ利用が急速に成長し、生データ記憶容量に対する需要が高まり続けている。企業が、電子商取引、オンライントランザクション処理、及びデータベースにますます依存するにつれて、管理及び記憶する必要がある情報量は莫大になる。その結果、記憶装置、より多くのユーザへのサービス提供、及び大量なデータのバックアップ作業を追加するための継続的な要求が作業を困難にするようになってきた。

**【0003】**

データに対するこのような需要の増大に対応するために、記憶領域ネットワーク（SAN）概念への人気が高まっている。SANはストレージネットワーキング・インダストリ・アソシエーション（SNIA）によって定義されたネットワークであり、その主目的はコンピュータシステムと記憶素子との間、及び記憶素子間のデータ伝送である。例えば、SCSI接続による記憶装置とサーバとの直接的な接続、及びイーサネット（登録商標）（例えば、NASシステム）等の従来型インタフェースによるLANへの記憶装置の増設とは異なり、SANは、直接接続型のSCSI及びNASと同様の帯域幅制限をもちにくい実質的に独立したネットワークを構築する。

**【0004】**

詳細には、SAN環境において、記憶装置（例えば、テープドライブ及びRAIDアレイ）及びサーバは、種々のスイッチ及び機器を介して相互接続されるのが一般的である。スイッチ及び機器との接続部は通常ファイバ・チャンネルである。一般的に、この構成により、SAN上の任意のサーバは任意の記憶装置と通信することができ、その逆も同様である。また、サーバから記憶装置への代替経路を提供する。換言すると、特定のサーバが低速であるか又は完全に利用不可能

である場合、SAN上の別のサーバが、記憶装置へのアクセスを提供することができる。また、SANは、データのミラーリングができるようにし、利用可能な複数のコピーを作り、結果的にデータ利用における信頼性をより高める。より多くの記憶装置を必要とする場合、特定のサーバと接続する必要なく追加の記憶装置をSANに増設することができる。それどころか、新しい記憶装置を記憶ネットワークに簡単に増設できるとともに、任意の地点からアクセスすることができる。

#### 【0005】

図1の機能ブロック図のシステム100にSANの一例を示す。図示したように、1又はそれ以上のサーバ102が存在する。例示的に3つのサーバ102のみが示されている。サーバ102は、イーサネット接続によってLAN106及び／又はルータ108に、そしてさらに、インターネット等のWAN110に接続される。さらに加えて、各々のサーバ102は、ファイバ・チャンネル接続によって、SANの「編地 (fabric)」と称されたりもする複数のファイバ・チャンネルスイッチ112の各々に接続される。例示的に2つのスイッチ112のみが示されている。次に、各々のスイッチ112は、複数のSAN機器114の各々に接続されている。例示的に2つの機器114のみが示されている。また、各々の機器は、テープドライブ、光学ドライブ、又はRAIDアレイ等の複数の記憶装置116の各々に結合される。更に、各々のスイッチ112及び機器114はゲートウェイ118に結合され、ゲートウェイ118はルータ108に結合され、最終的にルータ108はインターネット等の広域ネットワーク (WAN) 118に接続される。図1は、スイッチ112、機器114、記憶装置116、及びゲートウェイ118を含むSAN119として考えられる構成の一例を示す。なお、他の構成も可能である。例えば、1つの機器は、全スイッチ数よりも少ないスイッチに接続することができる。

#### 【0006】

機器114はSANの記憶管理を行う。機器114は、データを受信すると機器内のメモリにそのデータを記憶する。次に、正しい記憶装置にそのデータを転送するために、(機器内の) プロセッサを用いて当該データを分析し且つ操作す



る。一般的には、この記憶及び転送処理によりデータアクセスが減速してしまう。

#### 【0007】

機器はいくつかのスイッチングを行うが、多数のサーバ（3つ以上）が存在する場合があるため、及び、各々の機器のポート数が少ない（通常は2つ又は4つ）ため、スイッチ112は多数のサーバを少ない機器に接続する必要がある。それにもかかわらず、スイッチ112には殆ど処理能力（intelligence）が組み込まれておらず、選択された機器114にデータを転送するだけである。

#### 【0008】

機器が有する制限の1つは、一般に機器のポート数が非常に少ない（例えば、2つのポートのみ）という事実にある。その結果、機器が利用できる帯域幅は制限される場合がある。機器にポートを増設することは可能ではあるが、一般に非常に経費がかかるどの1又は2ポートも、高価なCPU又はサーバカードによってサポートされる。そこで一般に、ポートを増設するためには、（仮想化、並びに記憶及び転送機能を行う）全ファイルカードを装置に付加しなければならず、非常に経費がかかってしまうのが一般的である。もしくは、機器をSANに単に増設することもできるが、やはり非常に割高になりがちである。

#### 【0009】

更に、通常は機器114内において、SANは「仮想化」として既知の機能を実行するのが一般である。1又はそれ以上の物理的記憶装置上の空間が特定のユーザに割り当てられた場合に仮想化が実行されるが、この空間の物理的位置はユーザには分からないままである。例えば、ユーザは、自社の「エンジニアリング記憶空間」ENGにアクセスする、即ち、ユーザが外付けディスクドライブにアクセス又は「参照」しようとする場合、仮想空間ENGにアクセス及び「参照」することができる。しかしながら、ENG空間は、複数の物理的記憶装置上に分散させることができ、又は単一の記憶装置上に断片化させることさえできる。つまり、サーバが仮想装置（例えば、ENG）及びブロック番号を要求する場合、機器は、要求された仮想装置と物理的に相関関係がある装置を特定して、それに応じてデータを導く必要がある。

## 【0010】

一般に、SANは装置を相互接続する単一のプロトコルを使用して構築される。ファイバ・チャンネルは最も一般的に使用されているが、同様にイーサネット接続も使用されている。しかしながら、両方のプロトコルの使用が望まれる場合、2つのプロトコル間である種の変換を行う必要がある。この場合、ファイバ・チャンネルSAN119は、一般にブリッジ121を経由してイーサネットSAN122に結合される。一方のプロトコルから他方のプロトコルに変換するために、ブリッジにより受信されたパケットはメモリに記憶される。パケットがメモリに一旦記憶されると、プロセッサは、パケットを操作して、一方のプロトコルのヘッダを取り除きそして他方のプロトコルのヘッダを組立て、この結果全く新しいパケットを作る。詳細には、図2を参照すると、ブリッジ121によって（1又はそれ以上のパケットから成る）要求が受信されると、例えば、ホストバスアダプタ（HBA）202はこの要求をファイバ・チャンネル接続204上で受信する。プロセッサ208が要求を分析して操作する準備ができるまで、即ち、送信プロトコルに基づいて要求を再構築するまで、全ての要求がメモリ206に記憶される。プロセッサ208によって前記要求が操作されると、この要求はネットワークインタフェースカード（NIC）210に送信され、その後にイーサネット接続212上に送出される。勿論、逆の場合（イーサネットからファイバ・チャンネルへ）も同じ処理を行うことができる。従って、プロトコル間の変換処理は、かなりのメモリ及びプロセッサリソースを必要とし、このことは、データ伝送の遅延を引き起こすだけでなく、金銭及び不動産の両面におけるシステムのコストアップにつながる。しかしながら、現在利用可能な唯一の選択肢は、プロトコルを別個の各ネットワーク上で孤立させ続けることである。

## 【0011】

ゲートウェイ118（図1）は、SANをWANに接続するだけでなく、2又はそれ以上のSANを相互に接続するために使用される場合が多い。通常、ゲートウェイは、種々のプロトコル変換を行うのではなく、むしろ、本技術分野では既知であるように、IPパケットのデータをカプセル化する。それでもなお、複数のSANが接続される場合には、各々の接続装置に対して固有アドレスが存在す

る必要がある。しかしながら、IPプロトコルはアドレス指定用に32ビットを備えるが、ファイバ・チャンネルプロトコルは24ビットのビット数のみを備えるにすぎない。従って、大半のSANはファイバ・チャンネルを使用することから、ゲートウェイの使用にもかかわらず拡張性が問題になり、インターネット上でのSANの使用を制限する場合がある。

#### 【0012】

SANは数年前に導入されたが、普及に際して相互接続性の問題、利用可能な技術の不足、及び高い実行コストが大きな障害になっている。例えば、既存のSANは、配備コストが高く、管理コストも高い。図1を再度参照すると、一般的に、各々のスイッチ、機器、及びゲートウェイは、異なるベンダから提供され、管理基準の不足からベンダ専用の管理ツールが横行している。その結果、SANを配備するには、複数のベンダから機器を購入する必要がある。また、図1に示すように、各々のスイッチ、機器、ゲートウェイ、記憶装置、サーバ、及びルータは、管理ステーション120として示されるような独自の管理をもつことになる。独立した物理的な管理ステーションが示されているが、この独立した管理は、単一のコンピュータ上の独立したベンダ専用ソフトウェアの形態である場合が多く、そのソフトウェアは相互に連絡を行わないことを理解されたい。その結果、SANの集中管理が行われず、管理のために多くの人々を要求する複数の管理ステーションが存在するのが普通であれば、その管理コストが高くなってしまう。

#### 【0013】

(関連出願の説明)

本出願は、2001年9月28日出願の米国仮特許出願番号60/325,704「記憶領域ネットワークのために記憶装置スイッチ」の優先権を主張するものであり、その開示内容は、引用により本明細書に組み込まれている。

また、本出願は、本出願と同時出願され、その開示内容が引用により本明細書に組み込まれている以下の出願に関連する。

米国出願番号10/051,321「記憶領域ネットワークのための記憶装置スイッチ」

米国出願番号10/051, 164「サーバ不要の記憶サービス」

米国出願番号10/051, 093「記憶システムにおけるパケット分類」

米国出願番号10/051, 396「記憶システムにおける仮想化」

米国出願番号10/051, 339「記憶ネットワークにおけるサービス実行品質」

米国出願番号10/050, 974「記憶ネットワークにおける記憶リソースのプール及び準備」

米国出願番号10/051, 053「記憶リソースにおけるロードバランシング」

#### 【0014】

##### (発明の概要)

本発明の実施形態による記憶装置スイッチは、配備が容易で集中管理が可能なSANの構築を可能にする拡張性の高いスイッチである。更に、この記憶装置スイッチは、世界的規模のインフラ基盤の配備も可能にするので、記憶装置等のSANのリソースを実質的に世界中のどこにでも配置することが可能になる。更に、本発明による記憶装置スイッチにより、例えば、iSCSI（イーサネット接続上で搬送される最も新しく導入されたプロトコル）又はファイバ・チャンネルの両方を含む多重プロトコルSANが、任意のデータパケットを「ワイヤ速度」で処理すること、即ち、単にスイッチング又は経路指定機能を実行するスイッチによって生ぜしめる待ち時間の増加を招くことなく処理を行うことが可能になり、本発明によるスイッチは高帯域幅を有する。一般的に、ワイヤ速度でデータを処理するために、本発明の実施形態による記憶装置スイッチは、従来とは異なり、パケットのバッファリングを行わない。従って、従来技術に比較して、本発明の実施形態によるアーキテクチャでは、パケットを処理するための所要時間が最小になる。

#### 【0015】

詳細には、本発明によるスイッチは、ワイヤ速度で仮想化及び変換サービスを行う。このようなワイヤ速度の処理を行うために、スイッチ・ラインカードの全ポートには「処理能力(intelligence)」が分散されている。更に、各々のライン

カードは、パケットを分類し且つデータパケットを制御パケットと分離することができる。また、処理能力 (intelligence) が分散されるおかげで、各々のラインカードは、データパケット上で必要な場合には、(仮想アドレスを物理的アドレスへ変換する) 仮想化及び (第1のプロトコルの受信プロトコルを第2のプロトコルの発信パケットへ変換する) プロトコル変換を行い、そしてユーザ又はサーバが仮想化又は変換の必要性を意識することなく又は関与することなく行うことができる。処理能力を分散させたので、従来のCPU又はサーバカードよりも安価な多数のラインカードを作ることができ、例えば、多数のポートに対応するための記憶装置スイッチの拡張が容易になる。

#### 【0016】

更に、各々のスイッチは、ミラーリング、低速リンク上でのミラーリング、スナップショット、仮想ターゲットクローン化 (複製)、第三者コピー、定期的なスナップショット及びバックアップ、並びにリストア等のサーバ不要の記憶サービスを提供できる。スイッチがこのようなサービスに関する要求を受信すると、サーバ又は管理ステーション等の任意の他の装置の支援なしでこのサービスを実行することができる。

本発明を、以下の図面を参照しながら例示的な特定の実施形態に関して説明する。

#### 【発明を実施するための最良の形態】

#### 【0017】

本発明による記憶装置スイッチを含むシステム300を図3に示す。図示するように、このシステムは、既存のシステムに比べて非常に単純化されている。1つの実施形態において、システム300は、複数のサーバ302を含む。例示的に、3つのサーバ302が示されているが、別の実施形態において更に多くの又は更に少ないサーバを使用することができる。また、図示されていないが、サーバはLANに結合することもできる。図示のように、各々のサーバ302は、記憶装置スイッチに接続されている。しかしながら、別の実施形態において、各々のサーバ302は、存在する記憶装置スイッチ304の全数よりも少ない数のスイッチに接続することができる。サーバとスイッチとの間に形成される接続は、

任意のプロトコルを利用することができるが、1つの実施形態において、接続は、ファイバ・チャンネル又はギガビットイーサネット（iSCSIプロトコルに従いパケットを伝送）のいずれかである。他の実施形態では、インテル社によって定義されたインフィニバンド・プロトコル、又は他のプロトコル若しくは接続を使用することができる。図示の実施形態において、各々のスイッチは、複数の記憶装置又はサブシステム306の各々に順次接続されている。しかし、他の実施形態において、各々のスイッチは、記憶装置又はサブシステム306の全数よりも少ない記憶装置又はサブシステムに接続することができる。記憶装置スイッチと記憶装置との間に形成される接続は、任意のプロトコルを利用することができるが、1つの実施形態において、接続は、ファイバ・チャンネル又はギガビットイーサネットのいずれかである。特定の実施形態において、1又はそれ以上のスイッチ304は、都市規模ネットワーク（MAN）、又はインターネット308等の広域ネットワーク（WAN）にそれぞれ結合されている。記憶装置スイッチとWANとの間で形成される接続は、大半の実施形態においてはインターネットプロトコル（IP）が使用されるのが一般的であろう。MAN/WAN308と直接接続されるように示されているが、他の実施形態では、スイッチ304とMAN/WAN308との間の媒介手段としてルータ（図示せず）を利用することができる。更に、それぞれの管理ステーション310は、各々の記憶装置スイッチ304、各々のサーバ302、及び各々の記憶装置306に接続されている。管理ステーションは異なるコンピュータとして示されているが、各種装置を管理するためのソフトウェアは、まとめて1つのコンピュータ上に存在してもよいことを理解されたい。

#### 【0018】

図4は、本発明によるシステムの他の実施形態を示す。この実施形態において、2つのSAN402、404が形成されており、各々は、本発明の実施形態による1又はそれ以上の記憶装置スイッチ304を使用する。SAN402及び404は、スイッチ304によってインターネット等のWAN308を経由して接続されている。接続は、任意の規格又はプロトコルとすることができるが、1つの実施形態において、パケット・オーバーSONET（PoS）又は10ギガビッ

トイーサネットである。

#### 【0019】

図5は、本発明によるシステムの更に別の実施形態を示し、スイッチ304は相互に直接結合されている。図3又は図4に示す実施形態のいずれかにおいて、1より多いスイッチが使用される場合、これらのスイッチは、図5に示すように結合することができる。

#### 【0020】

本発明による記憶装置スイッチは、世界的に分散された共有記憶プールとして使用可能な記憶装置の集中管理を可能にし、世界的に分散された莫大な量の管理ステーション及び大勢の熟練した管理要員を有する代わりとなる。このような記憶装置スイッチは、「処理能力のある (intelligence)」スイッチであり、図3を図1と比較すれば分かるように、スイッチ、機器、及びゲートウェイの機能は、本発明の実施形態による記憶装置スイッチ304内に有効に統合されている。このような記憶装置スイッチ304は、スイッチング機能に加えて、一般的に従来のアーキテクチャの機器によってもたらされる仮想化及び記憶サービス（例えば、ミラーリング）を提供すると共に、プロトコル変換も提供する。また、本発明の特定の実施形態による記憶装置スイッチは、付加的な機能（例えば、仮想プライベートネットワークによるデータセキュリティ）を実行する。このような付加的機能としては、サーバによって従来より行われている負荷バランス (load balancing) 等の従来型システムにおける他の装置による機能、並びに従来型システムではこれまで利用不可能であった他の機能を挙げることができる。

#### 【0021】

本発明の実施形態による記憶装置スイッチの処理能力 (intelligence) は、全てのスイッチポートに分散される。この分散された処理能力は、システムの拡張性及び利用可能性を認めている。

#### 【0022】

更に、処理能力を分散したので、本発明の実施形態によるスイッチは、「ワイヤ速度」でデータを処理することができる。このことは、記憶装置スイッチ304が、一般的なネットワークスイッチ（図1のスイッチ112等）により生ぜしめ

るのと同程度の待ち時間をデータパケットにもたらしことを意味する。つまり、スイッチの「ワイヤ速度」は、特定ポートへの接続によって評価される。したがって、OC-48接続を有する1つの実施形態において、記憶装置スイッチは、OC-48速度（2.5ビット／ナノ秒）に遅れずについていくことができる。OC-48速度で送出される（10ビット／バイトによる）2キロバイトのパケットがスイッチに到来する所要時間は僅か8マイクロ秒である。1キロバイトのパケットの所要時間は僅か4マイクロ秒である。100バイトの最小パケットは、ほんの400ナノ秒しか要しない。しかし、本明細書において用語「ワイヤ速度」処理が用いられる場合は、この処理が100バイトのパケットを処理するのに僅か400ナノ秒であることを意味していない。1つの実施形態において、記憶装置スイッチは、OC-48速度、即ち約6マイクロ秒（4マイクロ秒／キロバイト又は2.5ビット／ナノ秒）で、（1バイトが10ビットになるように10ビット符号化した場合）1500バイトの最大イーサネットパケット、を処理できることを意味する。処理が1ビット／ナノ秒として定義されることが一般的な1Gbイーサネットポートを有する実施形態において、当該ポートの「ワイヤ速度」は10マイクロ秒／キロバイトになり、これはスイッチがキロバイトを処理するために最大10マイクロ秒であることを意味する。2Gbファイバ・チャンネルポートを有する実施形態では、「ワイヤ速度」が5マイクロ秒／キロバイトになる。更に別の実施形態では、10ギガビットイーサネット若しくはOC-192速度、又はそれ以上でデータを処理することができる。

### 【0023】

本明細書で用いる場合、「仮想化」は、ユーザが承認（subscribe）した仮想ターゲット空間を1又はそれ以上の物理的な記憶ターゲット装置上の空間へマッピングすることを実質的には意味する。「仮想」及び「仮想ターゲット」という用語は、承認毎に割り当てられた記憶空間が、記憶装置スイッチ304に接続する1又はそれ以上の物理的記憶ターゲット上のどこかに存在し得ることに由来する。物理的空間は、1又はそれ以上の「論理ユニット」（LU）を含み得る「仮想ターゲット」として提供することができる。各々の仮想ターゲットは、1又はそれ以上のLU番号（LUN）で特定される1又はそれ以上のLUから成り、i



SCSIプロトコル及びFCプロトコルで使用される場合が多い。各々の論理ユニット、従って各々の仮想ターゲットは、1又はそれ以上の領域、即ち物理的装置上の記憶空間の連続部分で構成されるのが一般的である。従って、仮想ターゲットは、記憶装置の全て（1つの領域）、単一の記憶装置の一部（1又はそれ以上の領域）、又は複数の記憶装置の一部（複数の領域）を占めることができる。物理的装置、LU、領域数、及びそれらの正確な位置は実体がなく、しかも承認ユーザに対して非表示である。

#### 【0024】

記憶空間は、多数の異なる物理的装置によってもたらされる場合があるが、各々の仮想ターゲットは、1又はそれ以上の領域（ドメイン）に属する。同じ領域のユーザのみが、その領域内の仮想ターゲットを共有することができる。領域集合によって、複数領域のユーザ管理が容易になる。領域集合に属するメンバは、同様に他の領域のメンバになることができる。しかし、本発明の実施形態において、仮想ターゲットは、1つの領域のみとすることができる。

#### 【0025】

図6は、本発明の実施形態による記憶装置スイッチ304の機能ブロック図を示す。1つの実施形態において、記憶装置スイッチ304は、複数のラインカード602、604、606、複数のファブリックカード608、及び2つのシステム制御カード610を含む。以下、それぞれについてその詳細を説明する。

#### 【0026】

##### システム制御カード：

2つのシステム制御カード（SCC）610の各々は、全てのラインカード602、604、606に接続する。1つの実施形態において、このような接続は、本技術分野で公知のI<sup>2</sup>C信号によって、SCCを備えたイーサネット接続によって形成される。SCCは、ファブリックカード（fabric card）と同様、I<sup>2</sup>C接続によって電源投入を制御し、且つ個々のラインカードをモニタする。また、イーサネット接続上のカード間通信を用いて、SCCは、以下に詳細に説明するスナップショット及び複製等の種々の記憶サービスを開始する。

#### 【0027】

更に、SCCは、サーバや記憶装置等のようなスイッチに取り付けられた全ての仮想ターゲット及び物理的装置の構成情報を追跡するデータベース612を保持する。

#### 【0028】

更に、データベースは、仮想ターゲット及びユーザについての異なる領域及び領域集合に関する情報と同様、使用、エラー、及びアクセスデータに関する情報を保持する。データベースの記録を、本明細書では「オブジェクト」と呼ぶ。各々のイニシエータ（例えば、サーバ）及びターゲット（例えば、記憶装置）は、世界中で唯一の識別子（World Wide Identifier : WWUI）を有し、それは本技術分野では既知である。データベースはSCC内のメモリ素子に保持され、1つの実施形態においてメモリ素子はフラッシュメモリから形成されるが、他のメモリ素子でも十分である。

#### 【0029】

管理ステーション（310）は、イーサネット接続を用いてSCC610を介して記憶装置スイッチ304に達することができる。したがって、SCCは管理ステーションと接続するための付加的なイーサネットポートを含む。管理ステーションの管理者は、SCCデータベース612に記憶された任意のオブジェクトに対する事実上の照会且つ更新と同様、記憶装置又は仮想ターゲットの追加又は取り外しを知ることができる。

#### 【0030】

2つのSCC610のうち、一方は主作動SCCであるが他方はバックアップ用であり、記憶装置内において作動が同期し続けるが、作動を直接制御するようにはなっていない。SCCは、高い利用モードで作動し、一方のSCCが故障した場合に他方は主制御装置になる。

#### 【0031】

##### ファブリックカード：

スイッチ304の1つの実施形態において、3つのファブリックカード608が存在するが、他の実施形態では、更に多くの又は更に少ないファブリックカードを有することができる。1つの実施形態において、各々のファブリックカード6

08は、ラインカード602、604、606の各々に結合されており、ラインカードの全てを相互に接続する働きをする。1つの実施形態においてファブリックカード608は、全てのラインカードが存在する場合は、最大トラフィックをそれぞれ処理することができる。各ラインカードによって処理されるこのトラフィック負荷は、1つの実施形態において最大160Gbpsであるが、他の実施形態では、これより多かったり少なかったりする最大トラフィック量を処理することができる。1つのファブリックカード608が故障すると、残りの2つのカードは、起こり得る最大スイッチトラフィックに対する十分な帯域幅をなおも有する。即ち、1つの実施形態において、各ラインカードは、入口側で10Gbps及び出口側で10Gbpsの20Gbpsのトラフィックを生成する。しかしながら、正常時には、3つのファブリックカードの全ては同時に有効である。各々のラインカードより、データに適合し得る3つのファブリックカードの任意の1つにこのデータトラフィックを送信する。

#### 【0032】

##### ラインカード：

ラインカードは、サーバ及び記憶装置への接続を形成する。1つの実施形態において、記憶装置スイッチ304は最大16枚のラインカードをサポートするが、他の実施形態では、異なる数のラインカードをサポートすることができる。更に、1つの実施形態において、異なる3種類のラインカード、即ち、ギガビットイーサネット（GigE）カード602、ファイバ・チャンネル（FC）カード604、及びWANカード606が利用される。他の実施形態では、これよりも多い又は少ない種類のラインカードを含むことができる。GigEカード602は、イーサネット接続用であり、1つの実施形態において、iSCSIサーバ又はiSCSI記憶装置（又は、他のイーサネット型装置）に接続する。FCカード604は、ファイバ・チャンネル接続用であり、ファイバ・チャンネルプロトコル（FCP）サーバ又はFCP記憶装置のいずれかに接続する。WANカード606は、MAN又はWANに接続するためのものである。

#### 【0033】

図7は、本発明による記憶装置スイッチ304の1つの実施形態において使用さ

れる汎用ラインカード700の機能ブロック図を示す。この図は、G i g E 6 0 2、F C 6 0 4、又はW A N 6 0 6等の全種類のラインカード間で共通の構成品を示す。他の実施形態において、他の種類のラインカードを利用して、インフィニバンド (Infiniband) 等の他のプロトコルを用いて装置に接続することができる。各ラインカードの相違点を以下に説明する。

#### 【0034】

##### ポート：

各ラインカード700は複数のポート702を含む。ポートはサーバ又は記憶装置のいずれかに対するラインカード接続を形成する。図示の実施形態では8つのポートが示されているが、他の実施形態では、これよりも多い又は少ないポートを使用することができる。例えば、1つの実施形態において、各々のG i g Eカードは、最大8つまでの1 G bイーサネットポートをサポートすることができ、各々のF Cカードは、最大8つまでの1 G bのF Cポート又は4つの2 G bのF Cポートのいずれかをサポートすることができる。そして各々のW A Nカードは、最大4つまでのO C - 4 8ポート又は2つのO C - 1 9 2ポートをサポートすることができる。従って、1つの実施形態において、接続可能な最大数は、スイッチ304毎に128ポートである。各々のラインカードのポートは全二重式であり、サーバ若しくは他のクライアントのいずれかに、又は記憶装置若しくはサブシステムに接続する。

#### 【0035】

更に、各々のポート702は関連メモリ703を有する。1つのポートに1つのメモリ素子だけが接続された状態で示されているが、各々のポートは、独自のメモリ素子を有することができること、又はポートの全ては、単一のメモリ素子に結合されてもよいことを理解されたい。説明を明瞭にする目的で、本明細書では1つのポートに1つのメモリ素子のみが接続された状態で示されている。

#### 【0036】

##### 記憶処理ユニット：

1つの実施形態において、各々のポートは、記憶処理ユニット (S P U) 701につながっている。S P Uは、データトラフィックを迅速に処理して、ワイヤ速

度の作動を可能にする。1つの実施形態において、SPUは、複数の構成要素、即ち、パケット統合及び分類エンジン(PACE)704、パケット処理ユニット(PPU)706、SRAM705、及びCAM707を含む。別の実施形態では、これよりも多い又は少ない構成要素を使用することができ、又は構成要素を組み合わせて同じ機能性を得ることができる。

#### 【0037】

##### PACE:

各々のポートは、パケット統合及び分類エンジン(PACE)704に結合されている。図示のように、PACE704は、2つのポートを統合して2倍の帯域幅をもつ単一のデータチャンネルにする。例えば、PACE704は、2つの1Gbポートを統合して単一の2Gbデータチャンネルにする。PACEは、以下に説明するように、受信した各々のパケットを制御パケット又はデータパケットに分類する。制御パケットは、ブリッジ716を経由してCPU714に送信され、処理される。データパケットは、ローカルヘッダが付加されて、以下に説明するパケット処理ユニット(PPU)706に送信される。1つの実施形態において、ローカルヘッダは16バイトであり、結果として、64バイト(ヘッダ16バイト、ペイロード48バイト)のデータ「セル」又は「ローカルパケット」になる。ローカルヘッダは情報を伝えるために使用され、そしてスイッチ204により内部で使用される。ローカルヘッダは、パケットがスイッチを出る前に取り除かれる。従って、本明細書で用いる場合、「セル」又は「ローカルパケット」は、スイッチ内で局所的に使用される送信単位であり、ローカルヘッダ及び元のパケットを含む(特定の実施形態において、元のTCP/IPヘッダも元のパケットから取り除かれる)。しかしながら、本発明の実施形態すべてが、ローカルヘッダを作成するか、又は外部パケットとは異なる「ローカルパケット」(セル)を有するものではない。したがって、本明細書で使用する場合、「パケット」という語は、「ローカル」パケット又は「外部」パケットを引用することができる。

#### 【0038】

分類機能は、スイッチが従来のシステムの記憶及び転送モデルを使用することな

く、記憶仮想化及びプロトコル変換機能をワイヤ速度で実行するのを助ける。各々のPACEは、PPU706への専用経路を有するが、図示の実施形態における4つのPACEの全ては、CPU714への経路を共有し、この経路は、1つの実施形態において104MHz/32(3.2Gbps)ビットデータ経路である。

#### 【0039】

##### パケット処理ユニット(PPU) :

PPU706は、作動中(on-the-fly)に仮想化及びプロトコル変換を行うが、このことは、本処理に関してセル(ローカルパケット)がバッファリングされないことを意味する。また、後述するスイッチ型記憶サービス機能を実行する。PPUは、1つの実施形態において、OC-48速度、即ち2.5Gbpsで入口側及び出口側の両方向にセルを移動させることができるが、他の実施形態において、OC-192速度、即ち10Gbpsでセルを移動させることができる。1つの実施形態のPPUは、入口側PPU706<sub>1</sub>及び出口側PPU706<sub>2</sub>を含み、両方を同時に実行する。入口側PPU706<sub>1</sub>は、PACE704から入力データを受信してデータをトラフィックマネージャ708へ送信するが、出口側PPU706<sub>2</sub>は、トラフィックマネージャ708からデータを受信してデータをPACE704へ送信する。

#### 【0040】

多数の記憶装置の接続(例えば、サーバから仮想ターゲットへ)は、各々のポートで同時に確立できる。しかしながら、各々の接続は、仮想ターゲットに固有のものであり、(iSCSI接続の場合)TCP制御ブロックインデックス及びポート番号によって一意的に特定することができる。接続が確立されると、ラインカード700のCPU714は、接続に関する仮想ターゲット記述子(VTD)をPPU706に送信することによって有効な仮想ターゲットを知らせる。VTDは、PPUがデータ(例えば、仮想化、変換、及び種々の記憶サービス)上で適切に作動するのに必要な接続及び仮想ターゲットに関する全ての関連情報を含む。VTDは、SCCデータベース内のオブジェクトから得られ、かつSCCデータベースの関連のオブジェクトに記憶されている情報の部分集合を含むのが普

通である。図7aは、本発明の1つの実施形態におけるVTDフィールドの一例を示す。しかしながら、本発明の他の実施形態は、これよりも多い又は少ないVTD、又はフィールドが異なるVTDを有することができる。

#### 【0041】

VTDを記憶して素早くアクセスすることができるように、1つの実施形態において、PPU706は、SRAM705及びCAM707に接続されている。SRAM705は、VTDデータベースを記憶する。また、VTD識別子(VTD ID)のリスト、即ち又はアドレスは、VTDに素早くアクセスするためにPPU CAM707に保持される。VTD IDは、TCP制御ブロックのインデックス及びLUNを使用して索引が付される(マッピングされる)。更に、IP経路指定サービスに関して、CAM707は、ルートテーブルを含み、CPUはルートが追加又は取り除かれた場合にルートテーブルを更新する。

#### 【0042】

1つのCAM及びSRAMのみが1つのPPUに接続されるように示されているが、これは説明を明瞭にするためであることに留意されたい。種々の実施形態において、各々のPPUは、それぞれのCAM及びSRAM装置に接続されることになるか、又はPPUの全てが単一のCAM及び／又はSRAMに接続されることになる。

#### 【0043】

PPUに対する各未処理の要求(例えば、読み出し又は書込み)に関して、要求状態を追跡するためにタスク制御ブロックがPPU SRAM707内に設けられる。入口側タスク制御ブロック(ITCB)は、入口側PPU上の記憶装置スイッチが受信した要求状態を追跡し、出口側タスク制御ブロック(ETCB)は、出口側PPU上の記憶装置スイッチが送信した要求状態を追跡する。各々の仮想ターゲット接続に関しては、多数の同時要求、従って、多数のタスク制御ブロックがあり得る。タスク制御ブロックは、要求開始時に割り当てられ、そして要求完了時に解放される。

#### 【0044】

トラフィックマネージャ:

各々のラインカード700上には、2つのトラフィックマネージャ(TM)708があり、1つは入口側トラフィックTMであり、1つは出口側トラフィックTMである。1つの実施形態において、入口側TMは、多重64バイトデータセルの形態で、パケットを4つのSPUの全てから受信する。この実施形態において、各々のデータセルは、16バイトのローカルヘッダ及び48バイトのペイロードを有する。ヘッダは、TMにセルの宛先ポートを教えるフローIDを含む。また、特定の実施形態において、SPUは、セルをTMに転送する前にTMヘッダをセルに付与することができる。また、TM又はSPUのいずれかは、特定の実施形態において、ファブリックカードを経由して送信できるようにセルを小さなセルに再分割することができる。

#### 【0045】

1つの実施形態において、入口側TMは、128ビット、104MHzインタフェース170を経由してデータセルをファブリックカードに送信する。出口側TMは、データセルをファブリックカードから受信して、それらを4つのSPUへ送る。

#### 【0046】

入口側TM及び出口側TMの両方は、送出のためのセル(ローカルパケット)を待ち行列に入れるための大きなバッファ712を有する。入口側TM及び出口側TM用の両バッファ712は64MBであり、多数のパケットを待ち行列に入れることができる。通常、SPUは、ファブリックカードの送信フローが受信フローと同じ速さなので、セルを入口側TMに素早く送信することができる。従って、セルは、出口側TMに素早く移動する。これに対して、発信ポートが詰まったり、又は複数の入口側ラインカードから送られたりすることから、出口側TMをバックアップすることができる。この場合、発信セルのヘッダにフラグを立てて、出口側SPUに迅速に対応策を取るように報知する。出口側TMは、フロー制御機能の起動要求を入口側SPUへ送信する。インターネット上の通信トラフィックとは異なり、記憶トラフィックに関しては、パケット落ちは許容できないことに注目することは価値がある。従って、バッファ内のセル量が所定の閾値を超えると、SPUは、バッファオーバーフローを回避するために、受信トラフィ



ックを減速するようフロー制御機能を直ちに作動させる必要がある。

#### 【0047】

##### ファブリック接続：

ファブリック接続710は、TMの256ビットパラレル信号（入口側128ビット、出口側128ビット）を、バックプレーン（backplane）に対して16ビットシリアルインタフェース（それぞれ入口側8ビット、出口側8ビット）に160Gbpsで変換する。従って、バックプレーンは、1/16のピンで実行中であるが速度は16倍である。この変換によって、何千ものピンやワイヤを接続することなく、適切なコストで高い利用性のあるバックプレーンを構築することができる。更に、1つの実施形態では3つのファブリックカードが存在するので、1つの実施形態において、各々のラインカード上に3つの高速コネクタがあり、各々のコネクタは、8ビット信号を3つのファブリックカードのそれぞれ1つに接続する。勿論、別の実施形態では、3つのファブリック接続710を必要としない場合もある。

#### 【0048】

##### CPU：

全てのラインカード上にはプロセッサ（CPU）714があり、プロセッサは、1つの実施形態において、PowerPC750Cxeである。1つの実施形態において、CPU714は、バスコントローラ715及びブリッジ716を経由して、3.2Gbバスで各々のPACEに接続する。更に、CPU714は、各々のPPU、CAM、及びTMに接続するが、特定の実施形態において、この接続は、40Mbpsの低速である。3.2Gb及び40Mb経路の両方によって、CPUは、ラインカード内の大半の素子と通信し、ラインカード上にある全素子の内部レジスタの読み出し及び書込みを行い、マイクロコードをダウンロードし、制御パケットを送受信することができる。

#### 【0049】

各々のラインカード上のCPUは、電源投入時に全てのチップを初期化し、マイクロコードが必要であればマイクロコードをSPU及び各々のポートに責任をもってダウンロードする。ラインカードが実行状態になると、CPUは、制御トラ

フィックを処理する。仮想ターゲット接続を確立するために必要な情報に関して、CPUは、SCCから情報を要求し、次に、これはSCCデータベース内の適切なオブジェクトから情報を取得する。

#### 【0050】

##### ラインカードとポートにおける区別：

1つの実施形態において、各種ラインカード内のポート、例えば、GigE、FC、又はWANは、各々のラインカードが一種類のポートのみをサポートするように区別できる。1つの実施形態の各種ポートについて以下に説明する。勿論、他のラインカードポートを、他の実施形態のインフィニバンド等の他のプロトコルをサポートするように設計することもできる。

#### 【0051】

##### GigEポート：

ギガビットイーサネットポートは、iSCSIサーバ及び記憶装置に接続する。GigEポートは、全種類のイーサネットトラフィックを伝送するが、本発明の1つの実施形態による記憶装置スイッチ304によってワイヤ速度で一般に処理される唯一のネットワークトラフィックは、TCP/IPパケット内部のiSCSIパケットデータユニット(PDU)である。しかしながら、他の実施形態において、イーサネット接続で伝送される他のプロトコル(ネットワークファイルシステム(NFS)等)によるパケットを、GigEポートで受信してSPU及び/又はCPUで処理することができる。

#### 【0052】

GigEポートは、仮想ターゲット又はiSCSI装置についてのTCP/IPセグメントを送受信する。仮想ターゲットに対するTCP接続を確立するために、ラインカードCPU714及びSCC610の両方が必要とされる。TCPパケットが受信され、初期接続手順が行われた後に、TCP制御ブロックが作成され且つGigEポートメモリ703に記憶される。また、接続を認証するとともに仮想ターゲットの構成を理解するために、SCCデータベースのオブジェクトからVTDを検索して、CPT SDRAM705に記憶する必要もある。TCP制御ブロックは、パケットが属する特定のTCPセッション又はiSCSI

接続を識別し、1つの実施形態において、TCPセグメント番号、状態、ウィンドウサイズ、及び場合によっては接続に関する別の情報を含む。更に、TCP制御ブロックは、本明細書では「TCP制御ブロックインデックス」として引用されるインデックスによって識別される。接続用VTDを作成して、SPU DRAM705に記憶する必要がある。CPUは、SDRAMに記憶され、当初はSCCデータベースから取得したVTD情報を検索することによってVTDを作成する。VTD IDは、VTDを素早く参照するためにSPU CAM707内のVTD IDリストに設定される。VTD IDは、TCP制御ブロックインデックスに関連付けられ、かつTCP制御ブロックインデックスによってインデックスが付けられる。

#### 【0053】

ポートはiSCSI PDUを受信すると実質的に接続の終端ポイントとして機能するが、その後、スイッチは、そのターゲットを用いて新しい接続を初期化する。入口側でパケットを受信した後に、ポートは、iSCSI PDUをTCP制御ブロックインデックスと共にPACEに送出して、特定のTCP接続を識別する。非TCPパケット又はiSCSI PDUを含まないTCPパケットに対しては、ポートは、接続の終端ポイントとしての機能を果たすことなく、パケットを送受信する。一般に、ポート702はPACE704と通信し、iSCSIパケットは、TCP制御ブロックインデックスを使用して送受信される。パケットのTCP制御ブロックインデックスが-1の場合は、非iSCSIパケットを特定する。

#### 【0054】

##### FCポート：

FCポートは、サーバ及びFC記憶装置に接続する。FCポートは、接続サーバにはファイバ・チャンネル記憶サブシステムとして見えるが、このことは、本技術分野では理解されているように、イニシエータが、接続を確立するためのプロセスログイン（PLOGI又はPRLI）が実行可能な仮想ターゲット装置の巨大プールのように見えることを意味する。FCポートは、GID拡張リンクサービス（ELS）を受け取り、そしてイニシエータ（例えば、サーバ）によるアク

セス可能なターゲット装置のリストを返す。

【0055】

ファイバ・チャンネル記憶装置に接続すると、ポートは、ファイバ・チャンネルFポートとして見えるが、このことは、本技術分野では理解されているように、記憶装置からファブリックログインを受け入れ、GID要求を受け入れ且つ処理することによってネームサービス機能を提供することを意味する。

【0056】

ポート初期化時に、ラインカードCPUは、ファブリックログイン、プロセスログイン、及びGIDの送受信を行う必要がある。SCCは、FC ELSをiSCSI要求及びiSCSI応答に変換するためのアプリケーションをサポートする。その結果、SCC内の同じデータベースは、iSCSIイニシエータ及びターゲットであるかのように、FCイニシエータ（例えば、サーバ）及びターゲット（例えば、記憶装置）を追跡し続ける。

【0057】

FC接続が確立すると、FCポートは、GigEポートとは異なり、TCP制御ブロック又はその等価物を作成する必要はない。全ての必要な情報は、FCヘッダから入手できる。しかし、(D\_IDによりインデックスされた)VTDは、依然として、GigEポートに関して説明したのと同様の方法で確立する必要があるであろう。

【0058】

FCポートは、1Gbポート又は2Gbポートとして構成することができる。1Gbポートとしては、図7に示すように、2つのポートが単一のPACEに接続されているが、2Gbポートとして構成される実施形態において、ポートトラフィック及びSPUが対応可能なトラフィックは、SPUでの混雑を避けるために一致させるべきである。1つの実施形態において、ポートは、POS/PHYインタフェースでPACEに接続する。各々のポートは、別個に構成することができ、即ち、1つのPACEは、2つの1Gbポートを有することができ、別のPACEは、単一の2Gbポートを有することができる。

【0059】

WANポート：

WANラインカードを含む実施形態において、1つの実施形態では、WANラインカードは、OC-48及びOC-192接続をサポートする。従って、2種類のWANポート、即ち、OC-48とOC-192とが存在する。OC-48に関して、各々のSPUに対して1つのポートがある。PACEには統合機能がないが、依然として分類機能はある。WANポートは、SONETに接続し、そしてICMP、RIP、BPG、IP、及びTCP等のネットワークパケットの送受信時にGigEポートのように機能する。GigEポートとは異なり、1つの実施形態におけるWANポートは、追加のハードウェア構成部品を必要とするVPN及びIPSecに対するネットワークセキュリティをサポートする。

**【0060】**

OC-192のワイヤ速度は高速になるので、OC-192をサポートする実施形態においては高速SPUが必要となるであろう。

**【0061】**スイッチ型の記憶操作：

本発明の1つの実施形態による記憶装置スイッチは、パケットの分類、仮想化、及び変換を含め、様々なスイッチ型の記憶操作を行なう。これらのサービスは、一般的にSPUによって行われる。1つの実施形態において、全てのポートはSPUを有し、制御トラフィックを処理するためのリソースを有するCPUへ制御トラフィックを送る間に、データトラフィックをできるだけ高速で処理することが可能になる。図7に示すように、4つのSPUは、8つのポートをサポートする単一のCPUを共有する。従って、データトラフィックは最小のリソース及びオーバーヘッドを使用し、各々がワイヤ速度で記憶トラフィックを処理するための処理能力を有する多数の低コストポートが可能になる。SPU機能については以下で詳細に説明する。

**【0062】**

SPU機能を説明する前に、iSCSI PDU（パケットデータユニット）及びFCIU（情報ユニット）の概要を説明することは好都合であろう。しかしながら、iSCSI及びFCプロトコルに関する一般知識をもっていることが前

提である。iSCSIに関する詳細は、引用により本明細書に組み込まれている、インターネットエンジニアリングタスクフォース（IETF）による続刊中のインターネットドラフト「draft-ietf-ips-iSCSI-07.txt」、2001年7月20日、を参照されたい。ファイバ・チャンネル（FC）に関する詳細は、引用により本明細書に組み込まれている、「情報システム—SCSI用dpANSファイバ・チャンネルプロトコル」改訂012、1995年12月4日（米国規格協会提案のドラフト）を参照されたい。

関連のPDU及びIUを、以下簡単に説明する。

### 【0063】

#### iSCSIコマンドPDU：

図8aはiSCSIコマンドPDUを示す。図示のように、以下のフィールドを有する48バイトを含む。第1のバイト（バイト0）において、XビットがイニシエータからターゲットまでのPDUについての再試行／再起動インジケータとして使用される。Iビットは、即時送出マーカとして使用される。Opcode 0x01は、iSCSI PDUの種類がコマンドであることを示す。また、バイト1は、複数のフラグ、F（最終）、R（読み出し）、及びW（書込み）を有する。また、バイト1は、通常3ビットであるタスク属性フィールドATTRを有する。バイト3のCRNは、SCSIコマンド参照番号である。AHS全長は、4バイト語における任意の追加の随機的なヘッダセグメント（図示せず）の全長を表す。データセグメント長は、ペイロードの長さを示す。LUNは、論理ユニット番号を指定する。イニシエータタスクタグは、タスクを識別するためにイニシエータ（例えば、記憶装置）によって割り当てられたタスクタグを識別する。期待データ伝送長は、操作に関するイニシエータに伝送されるか、又はそこから伝送されるデータのバイト数を示す。ExpStatSNは期待状態シーケンス番号であり、そしてExpDataSNは、期待データシーケンス番号である。コマンド記述子ブロック（CDB）は、一般に16バイトであり、SCSIコマンド自体を具体化する。

### 【0064】

#### iSCSI R2T PDU：

図8bはiSCSI R2T PDUを示す。バイト0において、0x31は、パケットをR2Tパケットとして識別する。イニシエータタスクタグは、コマンドPDUの場合と同じある。ターゲット伝送タグは、ターゲット（例えば、記憶装置）によって割り当てられ、データパケットの識別を可能にする。StatSNフィールドは、状態シーケンス番号を含む。ExpCmdSNは、次の期待CmdSNをイニシエータから識別し、MaxCmdSNは、イニシエータから受け入れ可能な最大CmdSNを識別する。R2TSNは、R2T PDU番号を識別する。期待データ伝送長は、ターゲットがイニシエータに送信してもらいたいバイト数を指定する（ターゲットは、複数のかたまり（chunk）でデータを要求できる）。従って、ターゲットはデータ伝送を開始すべきポイントを示すバッファオフセットも指定する。

#### 【0065】

##### iSCSIデータ書込み及び読み出しPDU：

図8cはiSCSI書込みデータPDUを示す。図8dはiSCSI読み出しデータPDUを示す。0バイトにおいて、0x05は、パケットを書込みパケットとして識別し、0x25は、パケットを読み出しパケットとして識別する。これらのPDUのフィールドの大半は、前述のPDUの場合と同じである。更に、DataSNは、データシーケンス番号を識別し、そして残余カウンタは例えば、イニシエータの期待データ伝送長が短すぎた場合に、期待されたバイトの何バイトが伝送されなかったかを識別する。

#### 【0066】

##### iSCSI応答PDU：

図8eはiSCSI応答PDUを示す。バイト0において、0x21は、パケットを応答パケットとして識別する。状態フィールドは、コマンドのSCSI状態を報告するために使用される。応答フィールドは、コマンドが完了したか、又は、エラー又は故障があったかを識別するiSCSIサービス応答コードを含む。基本の残余カウンタは、例えば、イニシエータの期待データ伝送長が短すぎた場合に、期待されたバイトの何バイトが伝送されなかったかを識別する。Bid\_\_読み出し残余カウンタは、期待されたバイトの何バイトがイニシエータに

伝送されなかったか示す。他のフィールドは、前述の他のPDUと同じである。

#### 【0067】

##### FCPフレームヘッダ：

各々のFCP情報ユニット(IU)は図8fに示すフレームヘッダを使用し、以下に説明するペイロードが続く。R\_CTLは、フレームをFC操作の一部として識別し、情報カテゴリーを識別する。D\_IDは、フレームの宛先を識別する。S\_IDは、フレームのソースを識別する。TYPEは、一般に、SCSI FCPシーケンスの全フレームについては0x80に設定される。F\_CTLは、シーケンス及び交換の開始、及び正常終了又は異常終了を管理する。SEQ\_IDは、特定の交換発信者と交換応答者との間の各々のシーケンスを固有値で識別する。DF\_CTLは、存在できる任意の随意的なヘッダを示す。SEQ\_CNTは、シーケンス内のフレーム順序を示す。OX\_IDフィールドは、交換発信者(イニシエータ)識別子である。RX\_IDフィールドは、交換応答者(ターゲット)識別子である。RLTV\_OFFフィールドは、情報カテゴリーの基本アドレスを基準にした各々のフレームのペイロードの最初のバイトの相対変位を示す。

#### 【0068】

##### FCP CMNDペイロード：

図8gはFCPコマンドIUのペイロードを示す。FCP\_LUNは、論理ユニット番号である。FCP\_CNTLは、複数の制御フラグ及びビットを含む制御フィールドである。FCP\_CDBは、アドレス指定された論理ユニットによって解釈されることになる実際のSCSI CDBを含む。FCP\_DLは、ターゲットに伝送されるか又はそこから伝送されることが期待されるデータバイトの最大数のカウントを含む。

#### 【0069】

##### FCP XFR\_RDYペイロード：

図8hはFCP XFR\_RDYのペイロードを示す。DATA\_ROフィールドは、次のFCP\_DATA IUの最初のデータバイトのRLTV\_OFFフィールドの内容を示す。BURST\_LENフィールドは、次のFCP\_DATA



A IUのために準備されたバッファ空間の大きさを示し、正確な長さのIUの伝送を要求する。

#### PCP DATA IU:

データIUに関するペイロードは、伝送された実際のデータである。

#### **【0070】**

#### FCP RSP IU:

図8 i はFCP応答IUのペイロードを示す。FCP\_\_STATUSフィールドは、コマンドタスクの正常終了時で0に設定される。正常に終了しなかった場合には、種々のステータス状況を示す。FCP\_\_RESIDフィールドは、このSCSIコマンドに対してFCP\_\_DATA IUにおいて伝送されなかった残余データバイト数のカウントを含む。FCP\_\_SNS\_\_LENは、FCP\_\_SNS\_\_INFOフィールドのバイト数を指定する。FCP\_\_RSP\_\_LENは、FCP\_\_RSP\_\_INFOフィールドのバイト数を指定する。FCP\_\_RSP\_\_INFOフィールドは、検出された任意のプロトコル障害を記述する情報を含む。FCP\_\_SNS\_\_INFOフィールドは、任意のセンスデータの存在を含む。

#### **【0071】**

各々のiSCSI PDU及びFC IUの詳細は、概括的に説明されている。iSCSI PDU、FC IU、及びこれらの各々のフィールドに関する更に詳細な説明は、前述のiSCSI文献及びFC文献に見出すことができる。

#### **【0072】**

#### 記憶装置スイッチの分類:

パケット又はフレーム（本明細書では全体として「パケット」と呼ぶ）が記憶装置スイッチに達すると、各々のポートにおいてデータ及び制御トラフィックに分離される。データトラフィックは、ワイヤ速度での仮想化及び変換のためにPPUに経路指定されるが、接続要求又は記憶管理要求等のデータトラフィックは、CPUに経路指定される。この分離を本明細書では「パケット分類」又は単に「分類」と呼び、一般にSPUのPACEで初期化される。従って、全パケットをCPUに送って処理する既存の技術とは異なり、本発明によるシステムは、データトラフィックを個別に高速で処理できるように、パケットコンテンツを認識

して、ワイヤ速度での処理を可能にすることを助長するようになっている。G i g E パケット及びF C フレームは、以下に説明するように、若干異なる方法で処理される。

#### 【0073】

G i g E ポートの入口側に到達するパケット（スイッチに到達するパケット）に関して、図9 a を参照しながら以下で各ステップを説明する。1つの実施形態において、G i g E ポートは、I P パケット又はi S C S I パケットのいずれかであるパケットを受信する（ステップ902）。パケットを受信すると、P A C E は、そのパケットとともに有効なT C P 制御ブロックインデックス（例えば、-1ではないインデックス）をポートから受信したか否かによって、仮想ターゲットアクセスが認識されたか否かを判定する（ステップ904）。有効なT C P 制御ブロックインデックスがある場合、次に、P A C E は、パケットのT C P ヘッダのフラグをチェックする（ステップ906）。T C P ヘッダのS Y N、F I N、及びR S T フラグが設定されている場合、C P U はT C P セッションの確立及び終了を行う責任があるので、パケットはC P U に送られる（ステップ916）。i S C S I T C P セッションが確立されると、T C P セッションを管理するために、G i g E ポートは、有効なT C P 制御ブロックをC P U から受信することになる。しかし、フラグが設定されていない場合、1つの実施形態において、P A C E は、T C P、I P、及びM A C ヘッダを取り除き（ステップ908）、i S C S I ヘッダを残し、次に、ローカルヘッダを追加する（ステップ910）。しかしながら、他の実施形態では、T C P、I P、及びM A C ヘッダを残して単にローカルヘッダを追加することができる。ローカルヘッダが追加されると、パケットはP P U に送信される（ステップ912）。

#### 【0074】

更に図10 a を参照すると、ステップ910が実行されると、受信されたT C P パケット1002は、ローカルパケット1004に変換され、I P、T C P、及びM A C ヘッダ1006、1008、1009が取り除かれ（1つの実施形態において）、ローカルヘッダ1010が追加される。しかしながら、場合によっては、i S C S I パケットに関するペイロードは、2つのT C P / I P パケットに

分割することもできる。従って、図10bを参照すると、受信されたTCPパケット1012は、ペイロードの第2の部分1014を含む場合もあり、ペイロードの第1の部分は先行パケットで送信済みである。ペイロードの第2の部分を含むパケットは、独立した新たなペイロード1016を追加的に含む。受信されたパケット1012は、2つのローカルパケット1018及び1020に分割される。ローカルパケット1018は、ローカルヘッダ1022及び先行パケットからのペイロード1024の第2の部分を含むが、iSCSIヘッダは含まない。ローカルパケット1020は、ローカルヘッダ1026、iSCSIヘッダ1028、及び新しいペイロード1030を含む。

#### 【0075】

図11は1つの実施形態において使用されるローカルヘッダ1100の一例を示す。ローカルヘッダ1100は、1つの実施形態において、以下のフィールドを含む。VTIDフィールドは、特定の接続に関するVTIDを識別するために使用される。フローIDは、パケットの宛先ポートを指定する。TCP制御ブロックインデックスは、特定の接続に関するTCP制御ブロックを指定する（TCP接続の場合）。TYPEフィールドは、データ又は制御といったパケット分類を指定する。サイズフィールドは、パケットサイズを示す。タスクインデックスは、スイッチ内のパケットを追跡して方向付けするために、並びに特定のタスクのパケットに関連する記憶情報の位置を見つけるために使用される。ローカルヘッダは、更に、ソース識別子（例えば、ソースポート、PACE、ラインカード、及び／又はCPUを識別する）及び宛先識別子（例えば、宛先ポート、PACEラインカード、及び／又はCPUを識別する）等の特定のハードウェア識別子を含む。

#### 【0076】

ローカルヘッダは、スイッチの全体にわたって、種々の装置（例えば、PACE、PPU）によって使用される。従って、ローカルヘッダの一部のフィールドを完全に使用できる場合もあり、フィールドコンテンツを交換又は更新できる場合もある。

#### 【0077】

再度図9 aを参照すると、有効なTCP制御ブロックインデックスがない場合（ステップ904）、パケットがIPパケットであるか否かが判定される（ステップ914）。パケットがIPパケットでない場合、CPUに送られる（ステップ916）。パケットがIPパケットである場合、次に、PACEは、宛先IPアドレスをチェックする（ステップ918）。IPアドレスが記憶装置スイッチのポートのIPアドレスと一致した場合、パケットは、CPUに送信され（ステップ916）、処理される。IPアドレスが記憶装置スイッチのポートのIPアドレスと一致しない場合、これは経路指定トラフィックであり、PPUに送られる（ステップ912）。

#### 【0078】

図9 bを参照すると、GigEポート宛のパケットが出口側でPACEによってPPU又はCPUから受信されると（ステップ950）、PACEは、ローカルヘッダを取り除く（ステップ952）。パケットがTCPセッション用である場合（ステップ954）、PACEは、GigEポートにその旨を知らせるために、ポートとのインタフェースの制御フラグを設定する（ステップ956）。パケットがTCPセッション用である場合、PACEは、インタフェース制御信号を使用してパケット及びTCP制御ブロックインデックスをポートに送る（ステップ958）。TCPセッションがない場合、パケットは単純にポートに送られる（ステップ960）。

#### 【0079】

図12 aは、FCポートから到着したパケットを分類する際にPACEにて行われるステップを示す。GigEポートの場合とは異なり、FCポートのPACEは、TCP制御ブロックインデックスを処理する必要はない。その代わりに、FCポートにてパケットを受信すると（ステップ1202）、FCPフレームヘッダのS\_IDフィールドは、フレームがオープンFC接続に属するか否かを判定するために参照されることができ、このステップは、パケットがPPUに送られた後に実行される。従って、PACEは、フレームがFCPフレームであるか否かだけを判定すればよく（ステップ1204）、このことは、フレームヘッダのR\_CTL及びTYPEフィールドを参照することで判定することができ

る。ローカルヘッダ1100（図11）が追加されるが（ステップ1206）、ヘッダ内のデータはPPUにとっては後で有用なので、FCPフレームヘッダはこの時点では取り除かない。次に、ローカルパケットはPPUに送られる（ステップ1208）。FCPフレームではない場合、フレームはCPUに送られる（ステップ1210）。

#### 【0080】

図12bを参照すると、FCポート宛のパケットが出口側でPACEによってPPU又はCPUから受信されると（ステップ1250）、PACEは、フレームをFCポートへ送る前に（ステップ1256）、単純にローカルヘッダを取り除く（ステップ1254）。ローカルヘッダは、PACEに対してパケットが（PACEが接続されている2つのポートのうちの）どのポート宛になっているのかを指示することになる。

#### 【0081】

GigEポート又はFCポートのいずれかで受信され、PPUに送られるパケットに関して、1つの実施形態において、PPUは、更に制御トラフィックを分離する。図13aを参照すると、PPUがパケットをPACEから受信すると（ステップ1302）、PPUは、IPパケットか又はTCPパケットかを判定する（ステップ1304）。パケットがIPパケットである場合、PPUは、CAMを検索してルートテーブルからパケットのフローIDを取得する（ステップ1306）。検索に失敗した場合、パケットは、未知の宛先IPアドレスを有し、これはCPUに送られ（ステップ1308）、CPUは、ICMPパケットをソースIPアドレスに逆送信する（ステップ1310）。検索によってフローIDが戻されると、パケットは、トラフィックマネージャに送られる（ステップ1311）。

#### 【0082】

受信されたパケットがTCPパケットである場合（ステップ1304）、PPUはTCP制御ブロックインデックスを使用してCAMを検索し、TCP制御ブロックインデックスは、TCPセッションを識別し、iSCSIヘッダからのLUNと一緒に、TCPセッションは、仮想ターゲットを識別して仮想ターゲット記

述子ID (VTD ID) を取得する (ステップ1312)。VTD IDは、実質的に、PPU SRAMに記憶されたVTDをアドレス指定するか又は指示する。PPUは、VTD IDを使用して、VTDのアドレスを取得するので (ステップ1312)、VTD IDの検索によって素早くVTDの位置を見つけることができる。VTDを取得できない場合にはiSCSIセッションがまだ確立されておらず、パケットはCPUに送られる (ステップ1314)。しかし、VTD IDがステップ1312で取得された場合、PPUは、パケットがiSCSI PDUを含むか否かを判定する (ステップ1315)。パケットがiSCSI PDUを含んでいない場合、パケットはCPUに送られる (ステップ1314)。しかし、パケットがiSCSI PDUを含む場合、PPUは、PDUがPDUを移動させるデータ (例えば、読み出しコマンド、書込みコマンド、R2T、書込みデータ、読み出しデータ、応答) であるか否かを判定する (ステップ1316)。PDUがPDU移動データでない場合、パケットは、CPUに送られる (ステップ1314)。しかし、PDUがPDU移動データである場合、以下に説明するように、PPUは、パケットに別の処理、例えば、仮想化及び変換処理を行なう (ステップ1318)。

#### 【0083】

PPUがFCPコマンドIUを有するFCPフレームを入口側で受信した場合、PPUは、図13aで説明するステップと同様のステップを実行する。即ち、ステップ1312においてCAM検索がVTD IDを見つけるためにFCPフレームからのS\_IDアドレス及びLUNを使用する点を除く、ステップ1302、1312～1318である。

#### 【0084】

図13bに示す出口側において、PPUは、トラフィックマネージャからパケットを受信した後 (ステップ1350)、ローカルヘッダのTYPEフィールドをチェックする (ステップ1352)。パケットがIPパケット又はCPU宛のパケットであることをフィールドが示す場合、PPUは、パケットをPACEに送信する (ステップ1354)。そうでない場合、以下に説明するように、PPUは、パケットに別の処理、例えば、仮想化及び変換処理を施す (ステップ1355)。

6)。

【0085】

前述のように、様々な状況において、SPUからCPUにパケットが送られることになる。この状況は以下を含む。

1. 宛先として記憶装置スイッチを有する非TCPパケット。このようなパケットは、本技術分野では理解されているように、ICMP、IP、RIP、BGP、又はARPパケットとすることができる。CPUは、スイッチ間通信及びIP経路指定機能を実行する。また、パケットは、SCCに送られることになるSLP又はiSNS要求であってもよい。
2. 適切な経路指定宛先に一致するCAMをもたないIPパケット。この状況は頻繁には発生しないはずであるが、発生した場合、CPUは、ICMPパケットをソースIPアドレスに戻す。
3. 非iSCSI TCPパケット。このパケットは、一般に、CPUがiSCSIに関するTCPセッションを確立又は終了させるためのものであり、典型的に、SYN、FIN、又はRSTフラグ集合を有するパケットである。
4. 非FCP FCフレーム。このフレームは、ネームサービスのためのFLOGI、PLOGI、及び他のFCP要求である。iSCSI TCPセッションと同様、これらのフレームによって、CPUは、FC装置を認識して通信することができる。1つの実施形態において、CPUは、SCCと通信してサービスを完了する必要がある。
5. SCSIコマンド、応答、又はデータではないiSCSI PDU。このパケットは、ピング (ping)、ログイン、ログアウト、又はタスク管理とすることができる。一般に、セッションが完全に確立される前に別のiSCSI通信が必要とされる。CPUは、ログインを完了するためにSCCデータベースからの情報を必要とすることになる。
6. 読み出し／書込み／検査ではないSCSIコマンドを有するiSCSIコマンドPDU。これらのコマンドは、仮想ターゲット動作が実行された場合にCPUによって処理されることになるiSCSI制御コマンドである。
7. 読み出し／書込み／検査ではないSCSIコマンドをもつFCPフレーム。

これらのコマンドは、仮想ターゲット動作が実行された場合にCPUによって処理されることになるFCP制御コマンドである。

#### 【0086】

##### 仮想化：

前述のようにパケットが分類された後に、PPUは、ワイヤ速度で仮想化を行ない、且つ1つの実施形態においてはデータバッファリングをせずに行なう。受信された各々のパケットに関して、PPUは、パケットの種類（例えば、コマンド、R2T/XFR\_DRY、書込みデータ、読み出しデータ、応答、タスク管理/停止）を判定し、次に、入口側アルゴリズム（パケットがスイッチに入る場合）又は出口側アルゴリズム（パケットがスイッチを出る場合）のいずれかを実行して、仮想ターゲットを物理的ターゲットに、又は物理的ターゲットを仮想ターゲットに変換する。従って、仮想化機能は、入口側ポートと出口側ポートとの間に分散される。別のワイヤ速度処理をできるようにするために、CAMと共に仮想記述子を使用して、要求位置をアクセス位置にマッピングするようになっている。更に、各々のパケットに関しては特別の配慮がなされてもよい。例えば、パケットの宛先である仮想ターゲットは、複数の非連続領域にわたって区切ってもよく、ミラーリングを行ってもよく、又はその両方を行ってもよい。（ミラーリングに関しては、本明細書の「記憶サービス」のセクションで説明される）。以下に各々のパケット種類に関する入口側プロセス及び出口側プロセスを説明する。しかしながら、一般的に、各々のパケットに関する入口側プロセスは、仮想ターゲットを確認し、パケットの宛先である出口側ポートを決定し、応答パケットを追跡できるように追跡タグを残す。一般的に、出口側プロセスは、追跡タグを保持し続け、且つブロックアドレスの調節を行い、仮想世界から物理的世界に変換する。

#### 【0087】

##### コマンドパケットー入口側：

仮想ターゲットへの又はそこからの伝送タスクを開始するために、SCSIコマンドは、それぞれiSCSI PDU又はFCP IUのiSCSI又はFCイニシエータによって常に送信される。図14及び図14aを参照すると、このパ



ケットが（分類後に）PPUで受信されると（ステップ1402）、次に、iSCSIイニシエータの場合にはTCP制御ブロックインデックス及び論理ユニット番号（LUN）を使用して、或いはFCイニシエータの場合にはS\_\_ID及びLUNを使用して、有効なVTD IDが存在するか否かを判定するためにPPU CAMをチェックする（ステップ1404）。各々の場合のLUNは、それぞれiSCSI PDU又はFCP IUで見つけることができる。有効なVTD IDが見つからない場合、応答パッケージがイニシエータに返送される（ステップ1406）。有効なVTD IDが見つかった場合、有効でないパッケージに関するチェックが行われる（ステップ1408）。このようなチェックとしては、仮想ターゲットに関する未処理コマンド番号が最大許容番号を超えたか否か、又はアクセス要求を受けたブロックが許容範囲内にあるか否かを判定するためのチェックを挙げることができる。無効パラメータが存在する場合、iSCSI又はFCイニシエータに応答パッケージが返送される（ステップ1406）。

#### 【0088】

チェックした全パラメータが有効な場合、図14aに示すように、タスクインデックスは入口側タスク制御ブロック（ITCB）と共に割り当てられる（ステップ1410）。タスクインデックスはITCBを指示又は識別する。ITCBは、（VTDから取得された）フローID、（iSCSIパッケージ自体からの）VTD ID、CmdSN、並びにiSCSI PDUに送信されたイニシエータタスクタグ、又はFCPフレームヘッダ内のOX\_\_IDを記憶する。ITCBは、PPU SRAMに記憶される。勿論、任意の所定時間に多数の処理中のコマンドが存在してもよいので、PPUは、任意の特定時間にITCB番号を記憶することができる。各々のITCBは、それぞれのタスクインデックスによって参照されることになる。

#### 【0089】

VTDは、特定の仮想ターゲットに対する未処理コマンドを追跡するので、新しいITCBが確立されると、未処理コマンド番号を増分する必要がある（ステップ1412）。特定の実施形態において、VTDは、特定の仮想ターゲットの任意の1つに対して未処理であろうコマンドの最大番号を設定する。フローID、

VTD ID、及びタスクインデックスの全ては、ローカルヘッダにコピーされる（ステップ1414）。フローIDは、トラフィックマネージャに宛先ラインカード及びポートを教える。その後、タスクインデックスは、パケットの特定のタスクを識別するために出口側ポートから返送されることになる。最後に、パケットは、トラフィックマネージャ、次に経路指定ファブリックへ送信されるので、最終的には出口側PPUに到達する（ステップ1416）。

#### 【0090】

仮想ターゲットが複数の領域で構成される場合、各々の領域に対して1つのVTDで識別される複数のフローIDが存在することになる。PPUは、パケットに関するブロックアドレスをチェックし、次に、正しいフローIDを選択する。例えば、仮想ターゲットが2つの1Gb領域を有し、コマンドのブロックアドレスが第2の領域にある場合、PPUは、第2の領域のフローIDを選択する。換言すると、フローIDは、宛先／出口ポートを決定する。読み出しコマンドが領域境界を越える場合には、コマンドが、第1の領域で開始ブロックアドレスを指定し、第2の領域で終了ブロックアドレスを指定することを意味し、適切なデータを第1の領域から読取った後、PPUは、残りのブロックを読取るために第2の領域にコマンドを繰り返す。領域境界を越える書込みコマンドに関しては、PPUは、両方の領域にコマンドをコピーすると共に書込みデータの順番を管理する。読み出しコマンドが領域境界を越える場合、2つの領域に対する2つの読み出しコマンドが存在することになる。第2の読み出しコマンドは、第1の読み出しコマンドの完了後にのみ送信され、データがイニシエータに連続的に確実に返送されるようにする。

図14aに関して、必ずしもローカルヘッダ内のフィールドの全てが図示されていないことに留意されたい。

#### 【0091】

##### コマンドパケットー出口側：

図15及び図15aを参照すると、出口側ポートに指定されたコマンドPDU又はIUは、スイッチファブリックを通過後に、PPUに到達する（ステップ1502）。次に、PPUは、パケットの宛先である物理的装置の識別を試みる（ス

テップ1504)。この識別を行うために、ローカルヘッダからのVTD IDを使用して、PTD ID (物理的ターゲット記述子識別子) に関するPPUCAMを検索する。VTD IDは、特定の出口側PPUに関連する特定のPTD IDに関係付けられ、インデックス付けされる。PTDは、VTDと同様に、PPUSRAMに記憶され、VTDで見出せるのと同様の情報を含む。検索が失敗した場合、これはCPUによって直接送信されたコマンドパケットであり、PPUによってどんな追加の処理も要求されないとみなされ、結果的に、PPUは、ローカルヘッダのフローIDに基づいてパケットを適切な出口側ポートに送る。検索が成功した場合、PTD IDは、仮想ターゲットがマッピングされ、かつパケットを現在処理中の特定の出口側ラインカードと通信状態にある物理的ターゲット (領域を含む) を識別する。

#### 【0092】

次に、図15aに示すように、PPUは、タスクインデックスを出口側タスク制御ブロック (ETCB) に割り当てる (ステップ1506)。1つの実施形態において、出口側に使用されるタスクインデックスは、入口側に使用されるものと同じである。また、タスクインデックスは、ETCBを識別する。更に、ETCBは、コマンドに必要な任意の他の制御情報を記憶するが、その制御情報としてはiSCSI PDUのCmdSN、又はFCPIUの交換シーケンスを挙げることができる。

#### 【0093】

次に、PTDコンテンツを用いて、PPUは、仮想ターゲットからのSCSIブロックアドレスを物理的装置のブロックアドレスに変換する (ステップ1508)。仮想ターゲットのブロックアドレスを領域の開始ブロックオフセットに加算することによって、この変換を行うことができる。例えば、アクセスしようとする仮想ターゲットブロックが1990であり、且つ対応する第1の領域の開始オフセットが3000である場合、アクセスされる領域のブロックアドレスは4990である。次に、PPUは、適切なiSCSI CmdSN又はFCPシーケンスIDを生成し (ステップ1510)、iSCSI PDU又はFCPフレームヘッダに入れる。また、PPUは、必要であればFCPフレームヘッダを構

築し（特定の実施形態において、入口側PPUは、FCPヘッダから必要な情報を読み出した後にFCPヘッダを取り除くが、他の実施形態では、FCPヘッダをそのままにしておき、必要なフィールドを単純に更新又は変更することができる）、又はiSCSIターゲットに送信されたパケットについては、TCP制御ブロックインデックスはPTDからローカルヘッダにコピーされる（ステップ1512）。更に、PPUは、iSCSI又はFCPヘッダに必要とされる任意のフラグ又は他の変数を与える。次に、完成したiSCSI PDU又はFCPフレームは、PACEに送信され（ステップ1514）、PACEは、ローカルヘッダを取り除き（ステップ1516）、適切なポートにパケットを送る（ステップ1518）。

#### 【0094】

複数領域の仮想ターゲットに関して、各々の領域は、異なる開始オフセットをもつ。従って、コマンドを2つの領域の間で分割する必要がある場合、PPUは、適切なアドレスを決める必要がある。例えば、仮想ターゲットが表1で定義した2つの領域をもつと想定する。

#### 【0095】

表1

領域	1	2
開始オフセット	3000	5000
ブロックサイズ	2000	2500

#### 【0096】

30個のブロックに関するアドレス1990で始まる仮想ターゲットにアクセスすることが望まれる場合、第1の領域のPPUは、コマンドを10個のブロックに関するアドレス4990に送る（5120バイトのデータ、1つの実施形態において、ブロックは512バイト）。第2の領域のPPUは、20個のブロックに関するコマンドをアドレス5000に送信する（10、240バイトのデータ）。換言すると、第1の領域のPPUは、アクセスされるアドレスを第1の領域の開始オフセットに加算し（3000+1990）、次に、そのアドレスを全サイズから減算して（2000-1990）、アクセスできるブロック数を決め

る。第2の領域のPPUは、開始オフセット(5000)で始まり、残りのブロック(20)を加算する(5000から5019)ことになる。別の実施例として、仮想ブロック2020にアクセスすることが望まれる場合、第2の領域のPPUは、第2の領域(5000)のオフセットを加算する前に、第1の領域のサイズ(2000)を減算し、結果的にアドレス5020を得ることになる。

#### 【0097】

##### R2T又はXFR\_RDY-入口側：

図16及び図16aを参照すると、前述のようにコマンドがターゲット記憶装置に送信された後、コマンドが書込みコマンドである場合、R2T PDU又はXFR\_RDY IUが、書込みデータを受け取る準備ができている場合に記憶装置から受信される(ステップ1602)。PPUは、イニシエータタスクタグ又はOX\_IDを使用することにより、対応するETCBを識別する(ステップ1604)。特定の実施形態において、パケットのイニシエータタスクタグ又はOX\_IDは、タスクインデックスと同じであり、ETCBを識別する。無効なイニシエータタスクタグ又はOX\_IDのために、PPUが有効なETCBを識別できなかった場合、パケットは破棄される。そうでない場合、ETCBを識別した状態で、PPUは、ETCBから入口側タスクインデックス(出口側タスクインデックスと異なる場合)及びVTD IDを検索する(ステップ1606)。また、PPUは、PTDからフローIDを検索し、フローIDは、ETCBにおいてPTD IDによって識別される。フローIDは、トラフィックマネージャに対して元のイニシエータ(入口側)ポートのラインカードを指示する。フローID、VTD ID、及びタスクインデックスは、パケットのローカルヘッダにコピーされる(ステップ1608)。最後に、パケットは、トラフィックマネージャ及びスイッチファブリックに送信される(ステップ1610)。

#### 【0098】

##### R2T又はXFR\_RDY-出口側：

R2T又はXFR\_RDYパケットがスイッチファブリックから出ていった後、イニシエータ(特定のタスクの元のコマンドを開始したデバイス)に返送される途中で、PPUによって受信される(ステップ1702)。タスクインデックス

は、PPUに対するITCBを識別し（ステップ1704）、ITCBから元のイニシエータタスクタグ及びVTD IDを取得することができる。R2T/XFR\_\_RDY期待データ伝送長、又はBURST\_\_LENフィールドは、ITCBに記憶される（ステップ1706）。ローカルヘッダは、FCP D\_\_ID又はTCP接続のためのTCP制御ブロックインデックスで更新される（ステップ1708）。ITCBに記憶されている、元の packets からの記憶されたS\_\_IDは、D\_\_IDになることに留意されたい。必要であれば、FCPフレームヘッダが構築されるか、又はそのフィールドが更新される（ステップ1710）。宛先ポート番号は、フローIDの代わりにローカルヘッダで指定され（ステップ1712）、イニシエータタスクタグと一緒にSCSI PDCに入れられるか、又は、FC接続の場合に、RX\_\_ID及びOX\_\_IDがFCPフレームに入れられる。また、PPUは、PDU又はFCPヘッダに入れる必要がある、他の任意のフラグ又は変数を入れる。パケットはPACEに送られ（ステップ1714）、PACEは、ローカルヘッダからの発信ポートを識別する。次に、ローカルヘッダは取り除かれ（ステップ1716）、送信に適したポートに送られる（ステップ1718）。

#### 【0099】

コマンドが2つ又はそれ以上の領域に分割される場合、例えば、コマンドが1つの領域で始まり別の領域で終わる場合、PPUは、第1の領域へのデータ伝送が完了するまで、第2の領域のR2T又はXFR\_\_RDYを保持する必要がある、これによってイニシエータからの連続的なデータ伝送を確実に行うことができる。更に、第2の領域のR2T又はXFR\_\_RDYのデータオフセットは、第1の領域に伝送されたデータ量を加算することによって変更されることが必要となるであろう。表1の実施例を参照すると、コマンドが30個のブロックに関するブロック1990にアクセスする場合、第2の領域のR2T又はXFR\_\_RDYのデータオフセットは、10個のブロックを追加する必要がある、その結果、11番目のブロックが第2の領域に伝送されることになる最初のブロックになる。

#### 【0100】

書込みデータパケットー入口側：

イニシエータは、R2T又はXFR\_RDYパケットの受信後に書込みデータパケットを返送する。図18及び図18aを参照すると、書込みデータiSCSI PDU又はFC IUがイニシエータから受信されると（ステップ1802）、パケットが属するITCBを識別する必要がある（ステップ1804）。通常、ITCBは、特定の実施形態においてタスクインデックスと同じである、RX\_ID又はターゲットタスクタグを使用して識別することができる。更に、SPUは、受信パケットが順序通りであることを識別する。しかしながら、特定の実施形態において、イニシエータは、要求していないデータ、即ち、R2T又はXFR\_RDYの受信前に送信されたデータを伝送することになる。この場合、PPUは、特定の仮想ターゲットの未処理タスクを検索することによってITCBを見つける必要がある。しかし、ITCBが見つからなかった場合、パケットは破棄される。ITCBが見つかった場合、伝送されることになるデータ総量がITCBにおいて更新される（ステップ1806）。フローID及びタスクインデックスは、パケットのローカルヘッダに追加される（ステップ1808）。次に、パケットは、トラフィックマネージャに送られ、最終的にはスイッチファブリックに送られる（ステップ1810）。

#### 【0101】

コマンドが2つの領域の間に分割される場合、コマンドは第1の領域で始まり第2の領域で終わるので、PPUは、特定のデータが属する領域を割り出して、データパケットを正しい出口側ラインカードに送る必要がある。PPUは、領域に対して正しいフローIDを設定する。第1の領域上のデータ伝送の完了後、PPUは、第2の領域のR2T又はXFR\_RDYを受信したか否かをチェックする。連続的な伝送を確実にを行うために、第1の領域上のデータ伝送が完了するまで、データは第2の領域に送信されないことになる。

#### 【0102】

書込みデータパケットー出口側：

図19及び図19aを参照すると、（トラフィックマネージャ経由で）スイッチファブリックから書込みデータパケットを受信すると（ステップ1902）、パケットのETCBを識別する必要がある（ステップ1904）。一般的に、ET

CBは、ローカルヘッダのタスクインデックスを使用して識別することができる。ETCBが識別されると、PPUは、ETCB内の情報を使用して、PDU又はFCPフレームヘッダに関するデータオフセット等の任意の別のフラグ及び変数と一緒に、適切なiSCSI Data SN又はFCPシーケンスIDを生成する（ステップ1906）。ローカルヘッダは、PTDからのTCP制御ブロックインデックス又はFCP D\_IDで更新される（ステップ1908）。また、ポート番号は、ローカルヘッダに追加される。完成したiSCSI PDU又はFCPフレームは、PACEに送信され（ステップ1910）、PACEは、ローカルヘッダを取り除き（ステップ1912）、パケットを適切なポートに送る（ステップ1914）。

### 【0103】

コマンドが2つの領域の間に分割される場合、第2の領域に対するパケットのデータオフセットが調整される必要がある。表1の実施例を使用すると、コマンドが30個のブロックに関する1990で始まる仮想アドレスにアクセスする必要がある場合、第2の領域に対する書込みデータパケットのデータオフセットは、実際にはイニシエータからの11番目のブロック11が第2の領域の最初のブロックなので、10個のブロック分を減算する必要がある。

### 【0104】

#### 読み出しデータパケットー入口側：

図20及び図20aを参照すると、読み出しコマンドの受信後に、ターゲット装置は、読み出しデータパケットに応答することになり、読み出しデータパケットは、（PACEでの分類後に）PPUにて受信されることになる（ステップ2002）。次に、パケットのETCBは、OX\_ID又はイニシエータタスクタグを使用して識別される（ステップ2004）。更に、PPUは、シーケンス番号を使用するか又はデータオフセットが昇順であることを検査することによって、パケットが順番に受信されたか否かを検査する（ステップ2006）。パケットが順序通りではない場合、読み出しコマンドはエラー終了する。しかしながら、パケットが適切な順番になっている場合、VTD ID、タスクインデックス、及びフローIDは、ETCB及びVTDから検索され、ローカルヘッダにコピー



される（ステップ2008）。パケットは、トラフィックマネージャに送信され、最終的にはスイッチファブリックに送信される（ステップ2010）。

#### 【0105】

読み出しデータパケットが領域境界を越える場合、第2の領域からのパケットのデータオフセットが変更される必要がある。通常、このオフセットは、フローIDが第2の領域からのパケットを識別することになるので、以下に説明するように出口側で行われる。更に、データを連続的に確実に返送するために、第2の領域への読み出しコマンドは、第1の領域からの読み出しが完了するまでは送信されないことになる。

#### 【0106】

読み出しデータパケットー出口側：

図21及び図21aを参照すると、PPUが読み出しデータパケットをスイッチファブリックから受信すると（ステップ2102）、パケットのITCBは、通常、ローカルヘッダのタスクインデックスを使用して識別される（ステップ2104）。ITCBから、PPUは、イニシエータタスクタグ又はOX\_IDを検索する（ステップ2106）。ITCBの保存データを使用して、PPUは、適切なiSCSI Data SN又はFCPシーケンスID、並びにPDU又はFCPフレームヘッダの他のフラグ又は変数を生成する（ステップ2108）。ローカルヘッダは、VTDからのTCP制御ブロックインデックス又はFCP\_S\_IDで更新される（ステップ2110）。しかしながら、イニシエータに返送されるパケットに関して、元のパケットからのS\_IDがD\_IDとして使用されることに留意されたい。また、発信ポート番号は、ローカルヘッダに追加される。次に、パケットは、PACEに送信され（ステップ2112）、PACEは、ローカルヘッダを取り除き（ステップ2114）、パケットを適切なポートに送る（ステップ2116）。

#### 【0107】

コマンドが2つの領域の間に分割される場合（ITCBで追跡した事実）、第2の領域からのパケットのデータオフセットは、前述の方法と同様の方法で変更する必要がある。

## 【0108】

応答パケットー入口側：

図22及び図22aを参照すると、応答パケットは、ターゲット装置から受信されることになる。次に、パケットに関するETCBは、パケットのイニシエータタスクタグ又はOX\_IDを使用して識別される（ステップ2204）。特定の実施形態において、イニシエータタスクタグ又はOX\_IDは、タスクインデックスと同じものとなる。ETCBが見つからなかった場合、パケットは破棄される。しかしながら、ETCBが見つかった場合、タスクインデックスは、VTD ID及びフローIDと一緒にパケットローカルヘッダにコピーされる（ステップ2206）。パケットは、トラフィックマネージャに送信され、最終的にはスイッチファブリックに送信される（ステップ2208）。最後に、応答パケットはタスクの完了を知らせるので、タスクのETCBは解放される（ステップ2210）。

## 【0109】

応答パケットー出口側：

図23及び図23aを参照すると、応答パケットは、スイッチファブリックを通過した後、出口側PPUによって受信されることになる（ステップ2302）。パケットに関するITCBは、ローカルヘッダからのタスクインデックスを使用して識別される（ステップ2304）。ITCBが見つからなかった場合、パケットは破棄される。ITCBが見つかった場合、仮想ターゲットに関する未処理コマンドカウントがVTDにおいて減分される（ステップ2306）。PPUは、LUN、iSCSI ExpStatSN、又はFCPシーケンスIDをITCB内の情報から生成し、必要であれば、適切なFCPヘッダを構築又は更新する（ステップ2308）。また、PPUは、PDU又はFCフレームヘッダのために他のフラグ及び変数を構築する。PPUは、VTDから検索されたTCP制御ブロックインデックス又はFCP S\_ID（これはD\_IDになる）を用いてローカルヘッダを更新する（ステップ2310）。パケットは、PACEに送られ（ステップ2312）、PACEは、ローカルヘッダを取り除き（ステップ2314）、パケットを適切なポートに送る（ステップ2316）。

PPUは、ITCBを解放する（ステップ2318）。

【0110】

書込みコマンドが2以上の領域に送信されている場合、応答パケットは、全領域に対する書込みが完了するまでイニシエータへ送信されない。

【0111】

図9から図23の全てについて、種々のステップが特定の順番で実行されるように説明されているが、他の実施形態において、特定のステップの順番を変更してもよく、特定のステップを同時に実行してもよいことに留意されたい。

【0112】

タスク管理PDU、異常終了、異常終了シーケンス／交換—入口側：

ABORT iSCSI機能、又は異常終了シーケンス／交換は、コマンドを通常とは異なる方法で終了する。PPUは、パケットのOX\_\_ID又はイニシエータタスクタグを使用してITCBを見つける。ITCBが見つからなかった場合、コマンドは、既に完了したか又は受信されなかったと想定され、TASK-NOT-FOUNDを示す応答が生成されることになる。ABORTがターゲット装置から受信された場合、PPUは、ETCBを見つけて解放する。ACKはターゲット装置に返送され、ABORTは、コマンドを終了するためにイニシエータに接続しているラインカードに送られる。ABORTがイニシエータから受信された場合、ABORTは、コマンドを終了するためにターゲットに接続しているラインカードに送られる。PPUは、それぞれのタスク制御ブロック、ITCB及びETCBを解放する。

【0113】

タスク管理PDU、異常終了、異常終了シーケンス／交換—出口側：

入口側ラインカードからのABORTは、出口側ラインカードに対して、ABORTをターゲット装置に送信することを指示する。完了応答がターゲットから返送される場合、ETCBは解放される。ETCBが見つからなかった場合、ABORTは無視される。

【0114】

変換：

前述のように、本発明による記憶装置スイッチは、複数のプロトコルのいずれかに基づいてデータを送信する装置に結合することができる。また、前述のように、1つの実施形態において、サーバ及び記憶装置が利用するプロトコルは、iSCSI及びファイバ・チャンネルである。しかしながら、スイッチが第1のプロトコルに基づいて作動するサーバに結合され、第2のプロトコルに基づいて作動する記憶装置に結合される場合、又は逆の場合も同様に、スイッチは、プロトコル変換を行う必要がある。従来、このような変換を行うためには、従来のシステムが仮にプロトコル変換を行うことができたとしても、パケットをメモリに記憶して転送前にCPUによって操作する必要がある。対照的に、本発明による記憶装置スイッチは、スイッチにおいてパケットのバッファリングを全く行うことなくプロトコル変換を行うことができる。

#### 【0115】

iSCSI PDU及びファイバ・チャンネル IUの両方は、それぞれのパケット又はフレームにSCSI CDB（コマンド記述子ブロック）を伝送するように設計される。従って、これらのプロトコルは、本発明の発明者が認識しているのと同様の意味論をもつ。表2は、各プロトコル間の比較を示す。

#### 【0116】

表2

SCSIフェーズ	iSCSIプロトコル	FCプロトコル
仲裁 (Arbitrate) 及び選択	イーサネットパケットを送る。	ファイバ・チャンネルフレームを送信する
コマンド	コマンドPDU	コマンドフレーム
切断 (Disconnect)	パケットを受信する	フレームを受信する
データ伝送のための再接続	R2T PDU	XFR_RDYフレーム
データ	TCPセグメント内のデータPDU	フレーム内のデータシーケンス
状態	応答PDU	応答フレーム
異常終了及びリセット	iSCSIタスク管理	ファイバ・チャンネル ELS
待ち行列満杯状態	Max Cmd SNウィンドウ	タスク設定満杯
セッションログインなし	iSCSIログイン及びログアウト	PLONI及びLOGO

#### 【0117】

表2から、iSCSIコマンドPDUとFCコマンドフレーム、R2T PDU

とXFR\_\_RDYフレーム、データPDUとデータフレーム、応答PDUと応答フレームとの間には相関関係があることが分かる。このような相関関係は、PPUにおいて、以下に説明するように、バッファリングを行うことなく1つのパケットから別のパケットへフィールドをマッピングすることで行われる直接的な変換に適する。異常終了及びリセット、セッションログイン及びログアウト、及び待ち行列満杯状態は、他のパケットに関連して不定期に起こり、ラインカードのCPUに送られて処理される（PPUによって実行される、SCSIデータ移動（読み出し／書込み）コマンドの異常終了を除く）。SCSIの仲裁及び選択、並びに切断に関して、iSCSI及びFCの両者は、単純にパケット／フレームを送受信することに留意されたい。

#### 【0118】

パケットがPPUへ到着すると、仮想化と同様に、PPUは、CAMを検索して受信コマンドが特定のセッション（iSCSI又はFCのいずれか）及び特定の仮想ターゲットに属するか否かを判定し、パケットに関連するVTDを識別する。CAM検索は、前述のように、TCP制御ブロックインデックス及びLUN（iSCSIパケットの場合）、又はS\_ID及びLUN（FCフレームの場合）を使用して行われる。しかしながら、本発明の1つの実施形態において、変換は、出口側PPU（スイッチファブリックを通過した後のパケットを受信するPPU）にて行われる。また、出口側PPUはCAMを検索するが、パケットのローカルヘッダのVTD IDを使用してPTDを見つける。

#### 【0119】

仮想化及び変換機能の両方に関して説明されているが、各種機能に関して説明した他のステップと同様に、CAM検索は、PPUによって1度だけ行われる必要があり、また、説明した全機能（例えば、分類、仮想化、及び変換）に対して行われる各種ステップは、多くの点で統合できる点に留意されたい。

#### 【0120】

同様に、仮想化機能に関して前述したように、VTDは、仮想ターゲット及び物理的ターゲットに関する変数を追跡し続けるが、同様に、PPUは、典型的にはITCB及びETCB（SCSIコマンド毎に各々1つ）の各プロトコル間で共

有されない変数を追跡し続ける。このような変数としては、iSCSIに関してはタスクタグ、CmdSN、DataSN、及びStatSNであり、ファイバ・チャンネルに関してはOX\_ID、RX\_ID、交換シーケンス番号、及びシーケンス開始フラグを挙げることができる。PPUが、VTD（又はPTD）並びにそれぞれのETCB又はITCBを有すると、変換を行うのに必要な全ての情報を有する。iSCSIからFCへの変換、又はその逆の変換は、一般的に、受信パケット（例えば、iSCSI）のフィールドから情報を取得すること、及び情報を送信パケット（例えば、FCP）の対応フィールドへマッピングすることとを必要とする。

### 【0121】

FCターゲットに対するiSCSIイニシエータ：

まず、iSCSIイニシエータ（サーバ）からFCターゲット（記憶装置）への変換について説明する。iSCSIコマンドPDUからFCP\_CMND IUへの変換は、以下の表3に基づいて行われる。また、図8a～図8iを参照されたい。

### 【0122】

表3

iSCSIコマンドPDUから	FCP_CMND IUへ
iSCSI PDUのLUNフィールド	FCP_LUN
ATTR (3ビット)	FCP_CNTL
CDBフィールド	FCP_CDB
期待データ伝送長	FCP_DL
	OX_ID、SEQ_ID、SEQ_CNT

### 【0123】

表3によれば、iSCSI PDUのLUNフィールドの内容は、FCP\_CMND IUのFCP\_LUNフィールドにマッピングされる。物理的ターゲットのLUNは、PTDから取得される。iSCSIタスク属性フィールドATTRの3ビットのみが、FCP\_CNTLフィールドにマッピングされる。iSCSI PDUのCDBフィールドの内容は、FCP\_CDBフィールドにマッピングされる。データ伝送サイズフィールドの内容は、FCP\_DLフィールドにマ

ッピングされる。OX\_IDは、FCPフレームヘッダに固有なので、ターゲットからの各々のパケットの識別を容易にするために、典型的にETCBからのタスクインデックスを用いて、PPUによって情報が与えられる。FCPフレームヘッダの他のフィールドは、PTD又はVTDからの情報を用いて簡単に生成することができる。

#### 【0124】

FC記憶装置が応答する場合、FC記憶装置は、FC XFR\_RDYフレームで応答することになり、FC XFR\_RDYフレームは、iSCSI R2T PDUに再変換する必要がある。

#### 【0125】

表4

FCP XFR_DRYから	R2T iSCSI PDUへ
DATA_RO	バッファオフセット
BURST_LEN	データ伝送長
	イニシエータタスクタグ及び他のフィールド

#### 【0126】

表4に示すように、バッファオフセット及びデータ伝送長フィールドは、FC XFR\_RDYフレームから直接マッピングされることができる。しかしながら、StatSN、ExpCmdSN、MaxCmdSN、及びR2TSN等の他のフィールドをITCBから取得する必要がある。更に、iSCSI R2T PDU固有のタスクタグのような変数は、通常、PTD又はVTDからのフィールドを使用して、PPUによってパケットに入れられる。

#### 【0127】

R2Tの受信後、iSCSI イニシエータは、書込みデータPDUを送信することになり、書込みデータPDUは、FCP Data IUに変換する必要がある。

#### 【0128】

表5

iSCSI読み出しデータPDUから	FCP DATA IUへ
バッファオフセット	RLVT_OFF
ペイロード	ペイロード
	OX_ID、SEQ_CNT

## 【0129】

表5に示すように、FCP DATA IUのRLVT\_OFFフィールドは、iSCSI PDUのバッファオフセットフィールドからマッピングされることになる。各々のパケット/フレームに関するペイロードは、全く同じである。更に、ETCBから取得した、OX\_ID及びSEQ\_CNT等のFCフレーム固有の変数が追加される。

## 【0130】

iSCSIイニシエータから最初に送信されたiSCSIコマンドが読み出しデータコマンドである場合、FCターゲットは、FCP\_DATA IUで応答することになり、FCP\_DATA IUは、iSCSI読み出しデータPDUに変換する必要がある。

表6

FCP DATA IUから	iSCSI読み出しデータPDUへ
RLVT_OFF	バッファオフセット
データペイロード	データペイロード
	イニシエータタスクタグ、残りカウント

## 【0131】

表6に示すように、iSCSI PDUのバッファオフセットは、FCP IUのRLVT\_OFFフィールドからマッピングされることになる。他の全フィールドは、ITCB並びにタスクタグ等のPDU固有の変数から取得される。

## 【0132】

タスクが完了すると（例えば、データの読み出し又は書込みが終了すると）、FCPターゲットは、iSCSIフォーマットに変換する必要がある応答パケット（FCP\_RSP IU）を送信する。

表7



FCP応答IUから	iSCSI応答PDUへ
FCP__STATUS	フラグ及び状態フィールド
FCP__SNS__LEN	データセグメント長
FCP__RESID	基本残余カウンタ
FCP__SNS__INFO	センスデータ
FCP__RSP__INFO	エラーコード
	イニシエータタスクタグ、MaxCmdSN、ExpCmdSN

## 【0133】

表7に示すように、FC IUの状態フィールドは、iSCSI PDUのフラグ及び状態フィールドにマッピングされる。FCP\_\_SNS\_\_LEN、FCP\_\_RESID、及びFCP\_\_SNS\_\_INFOは、それぞれ、データセグメント長、基本残余カウンタ、及びセンスデータにマッピングされる。FCP\_\_RSP\_\_INFOフィールドは、iSCSIエラーコードにマッピングされる必要がある伝送エラー用である。最後に、iSCSI状態PDUに固有のタスクタグ、又はExpCmdSN、StatSN、MaxCmdSN、ExpDataSN、ExpR2TSN等の変数はITCB又はVTDから追加される。

## 【0134】

異常終了タスクセット等のタスク管理用のFCP\_\_CNTLにフラグがある場合、別個のiSCSIタスク管理コマンドは、iSCSIイニシエータ装置に送信されることになる。同様に、iSCSIタスク管理PDUが受信された場合、FCP\_\_CNTLに適切なフラグを有するNOP PFコマンドがターゲット装置に送信される。

## 【0135】

前記の表には、iSCSI PDU又はFCPフレームのいずれかに固有の全フィールドが記載されていない点に留意されたい。フィールドを完全に記載するために図8aから図8iを参照することができる。記載されていない任意のフィールドについては、関連のタスク制御ブロック、VTD、PTDから取得することができ、又は簡単に生成することができることを理解されたい（例えば、FCP形式フィールドは、常に0x08である）。

## 【0136】

iSCSIターゲットに対するFCイニシエータ：

FCPからiSCSIへの変換は、iSCSIからFCPへの変換の逆である。  
 この場合も変換は出口側PPUで行われる。最初に、FCPイニシエータはFC  
 Pコマンドを送信することになるが、FCPコマンドは、iSCSIターゲット  
 に適するように変換する必要がある。

## 【0137】

表8

FCPコマンドIUから	iSCSIコマンドPDUへ
FCP__LUN	LUN
FCP__CNTL	ATTR
FCP__CDB	CDB
FCP__DL	期待データ伝送長
	CmdSN、タスクタグ、ExpStatSN

## 【0138】

表8に示すように、FC IUのLUN、CNTL、CDB、及びDLフィールド  
 は、iSCSI PDUのLUN、ATTR、CDB、及びデータ伝送サイズ  
 フィールドにマッピングされる。更に、CmdSN及びタスクタグ等のiSCS  
 I PDU固有の変数は、PPUによって作成され、CmdSN及びタスクタグ  
 の両方は、ETCBから取得することができる。データセグメント長フィール  
 ドは、FCPフレームに関する当面のデータがないためにゼロになる。

## 【0139】

iSCSIターゲットがコマンドを受信した後（コマンドは書込みコマンド）、  
 ターゲットは、R2T PDUで応答することになり、R2T PDUは、FC  
 P XFR\_\_RDY IUに変換する必要がある。

## 【0140】

表9

iSCSI R2T PDUから	FCP XFR__RDY IUへ
バッファオフセット	DATA__RO
データ伝送長	BURST__LEN
	RX__ID、SEQ__ID

## 【0141】

表9に示すように、iSCSI PDUのバッファオフセット及びデータ伝送長フィールドは、XFR\_RDY IUのDATA\_RO及びBURST\_LENフィールドにマッピングされる。更に、PPUは、ITCBで利用可能なRX\_ID及びSEQ\_ID等のFCP IUに固有の変数を追加する。

## 【0142】

FCイニシエータは、XFR\_RDY IUの受信後に、iSCSIフォーマットに変換する必要がある書込みデータを送信することになる。

## 【0143】

表10

FCPデータIUから	iSCSI書込みデータPDUへ
RLVT_OFF	バッファオフセット
ペイロード	ペイロード
	Data SN、ExpCmd SN、ターゲットタスクタグ

## 【0144】

表10に示すように、書込みデータに関して、FCP IUのRLVT\_OFFは、iSCSI PDUのバッファオフセットにマッピングされるが、各々のペイロードは同じである。更に、他のフィールドは、iSCSIデータPDUに固有のData SN等の変数も含めてETCBから取得される。

## 【0145】

元のイニシエータコマンドが読み出しコマンドである場合、iSCSIターゲットは、FCPフォーマットで入れる必要がある読み出しデータを用いて応答することになる。

## 【0146】

表11

iSCSI読み出しデータPDUから	FCP DATA IUへ
バッファオフセット	RLVT_OFF
ペイロード	ペイロード
	RX_ID、SEQ_ID

## 【0147】

表11に示すように、バッファオフセットフィールドは、FCP IUのRLVT\_\_OFFフィールドにマッピングされるが、両者のペイロードは同じである。更に、PPUは、ITCBに見つけることができる、RX\_\_ID及びSEQ\_\_ID等のFCP IUに固有の変数を追加する必要がある。

## 【0148】

最後に、タスクが完了すると、iSCSIターゲットは、応答PDUを送信することになり、応答PDUは、FCP RSP IUに変換する必要がある。

## 【0149】

表12

iSCSI 応答PDUから	FCP RSP IUへ
フラグ及び状態	FCP__STATUS
データセグメント長	FCP__SNS__LEN
基本残余カウンタ	FCP__RESID
センスデータ	FCP__SNS__INFO
伝送エラー	FCP__RSP__INFO
	OX__ID、SEQ__ID

## 【0150】

表12に示すように、iSCSI PDUのフラグ及び状態は、FCP IUのSTATUSフィールドにマッピングされる。iSCSIフィールドのデータセグメント長、基本残余カウンタ、及びセンスデータの全ては、FCP IUのFCP\_\_SNS\_\_LEN、FCP\_\_RESID、及びFCP\_\_RSP\_\_INFOフィールドにそれぞれマッピングされる。伝送エラーは、FCP IUのFCP\_\_RSP\_\_INFOフィールドにマッピングされる。更に、OX\_\_ID及びSEQ\_\_ID等のFCP IUに固有の変数は、PPUによって追加される。

## 【0151】

異常終了タスクセット等のiSCSIタスク管理パケットを受信した場合、これはFCP\_\_CNTLフィールドにタスク管理フラグをもつNOP コマンドを使用してFC装置に送信されることになる。

## 【0152】

前記の表には、iSCSI PDU又はFCPフレームのいずれかに固有の全フ

フィールドが記載されていない点に留意されたい。フィールドを完全に記載するために図8a～図8iを参照することができる。記載されていない任意のフィールドについては、関連のタスク制御ブロック、VTD、PTDから取得することができ、又は簡単に生成することができることを理解されたい（例えば、FCP形式フィールドは、常に0x08である）。

### 【0153】

#### 記憶サービス：

本発明の実施形態によるスイッチは、タスクを複数のラインカード上に分散することによって、スイッチ型の記憶サービスをワイヤ速度で行うので、スループットを最大にすることができる。本発明の1つの実施形態で提供される記憶サービスとしては、ローカルミラーリング、低速リンク上でのミラーリング、スナップショット、仮想ターゲットクローニング（複製）、第三者コピー、定期的スナップショット及びバックアップ、及びリストア等を挙げることができる。これらのサービスの各々について以下で更に詳細に説明する。他の実施形態では、これよりも多い又は少ないサービスを行うことができる。

### 【0154】

特定のサービスを説明する前に、図24を参照すると、概括的に、記憶サービスは、最初に記憶装置スイッチとのイーサネット接続上の管理ステーション（又は他の装置）によって起動される（ステップ2402）。このようなイーサネット通信は、1つの実施形態においてSCC610（図6）で行われる。SCCは、データベースを通じて、そのサービスのためのラインカードを判定し、そしてVTD及びLUN情報を含む全ての関連情報をこれらのラインカードに送り、そのサービスを実行する（ステップ2404）。SCCがラインカード毎に有する全ての情報は、イーサネット通信上でのカード間通信を使用して、SCCからラインカードへ送られる。次に、ラインカードは、要求された実際のサービスを行う（ステップ2406）。タスクが完了すると、SCCは、管理ステーションへの返送応答を開始して（ステップ2408）、サービスの完了を指示する。従って、従来のシステムとは異なり、管理ステーションは、サービス要求を開始する以外は、そのサービスに関与する必要は全くない。

## 【0155】

ローカルミラーリング：

仮想ターゲットがミラーリングされた場合、即ち、そのデータと全く同じコピーが2つの別個の物理的位置に記憶された場合、ミラーリングされた仮想ターゲットの「メンバ」と呼ばれる場合が多い。VTD内のフローIDは、パケットが複数の出口側ポートにマルチキャストされることを表す。ミラーリングされた仮想ターゲットにおいて、書込みコマンドが領域境界を越える場合、PPUは、ミラーリングされたターゲットの各々のメンバに対する各々の領域に関するパケットを複製することになる。また、PPUは、適切なフローIDをトラフィックマネージャに送り、トラフィックマネージャは、受信した各々のコマンドを複数の出口側ポートへ送信する。ミラーリングされた仮想ターゲットからの読み出し時に、PPUは、最小平均応答時間を有するミラーリングされたターゲットの1つのメンバを選択する。そのメンバのフローIDは、読み出しコマンドを選択された出口側ポートに導く。応答時間は、VTDで利用可能である。

## 【0156】

書込みコマンドの送信後に、ミラーリングされたターゲットのメンバの1つからR2T又はXFR\_RDYを受信した場合、PPUは、全てのメンバ及び／又は領域がR2T又はXFR\_RDYを返送するまで待機する。全てのメンバが応答済みになると、PPUは、データを受信するために利用可能な最小ブロックを指定するR2T又はXFR\_RDYを、イニシエータに送信する準備を行うことになる。即ち、データが返送されると、データは全てのミラーリングされたメンバにマルチキャストされることになるが、メンバは、要求した以上のデータを受信することはできない。従って、PPUは、ITCBにおいて、各々の領域に関するR2T又はXFR\_RDYで指定された要求データ量を追跡する必要もある。最小量のデータが受信され（イニシエータから）、ミラーリングされたターゲットの各々のメンバにマルチキャストされると、PPUは、最小量のデータを要求した領域が別のR2T又はXFR\_RDYを送信するのを待つ。2つの（又はそれ以上の）ターゲットが最小量のデータを要求した場合（即ち、両者が同じデータ量を要求した場合）、PPUは、最小量を要求した両方の（又は全ての）ター

ゲットが別のR2T又はXFR\_RDYを送信するまで待機する。次に、PPUは、全ての領域の最小残量のR2T又はXFR\_RDYを返送する。このプロセスは、全領域が全ての要求データをもつまで続く。1つの実施例を表13に示す。

【0157】

表13

	領域1	領域2	イニシエータへ
書込まれる全データ	4 k	4 k	
第1のR2T又はXFR_RDYで指定されたサイズ	2 k	3 k	
イニシエータからのPPU要求			2 k
不満足なR2T又はXFR_RDY (2 k 書込み後)	0 k	1 k	
第2のR2T又はXFR_RDYで指定されたサイズ	2 k		
イニシエータからのPPU要求			1 k
不満足なR2T又はXFR_RDY (1 k 書込み後)	1 k	0 k	
第3のR2T又はXFR_RDYで指定されたサイズ		1 k	
イニシエータからのPPU要求			1 k
不満足なR2T又はXFR_RDY (1 k 書込み後)	0 k	0 k	

【0158】

#### 低速リンク上のリモートミラーリング：

前述のように、ミラーリングは、2つの同じデータ集合の各々が、別個の物理的位置にそれぞれ記憶される場合に起こる。大部分の従来システムは、ローカルミラーリング、即ち、同一SAN上に存在する各装置におけるミラーリングだけをサポートする。しかしながら、本発明の実施形態は、低速リンク上でのミラーリングをサポートする。例えば、データの1つのコピーが1つのSAN上にあり、データの第2のコピーがSANから離れた位置に、例えば、第2のSAN上に記憶されている場合のミラーリングをサポートする。例えば、図4を参照すると、データのローカルコピーがSAN402にあるが、リモートミラーコピーは、SAN404にあることができる。従って、リモートミラーリングは、本発明の実施形態のスイッチにおいて可能になり、インターネット等のWANを介してデー

タをターゲットへエクスポート（又はインポート）することができる。

【0159】

しかしながら、低速リンク上のミラーリングとローカルミラーリングとの間の1つの重要な相違点は、リモートターゲットとの通信における固有の待ち時間である。例えば、WAN上でリモートターゲットと通信する場合の平均待ち時間は8  $\mu$ s/マイルである。従って、リモートターゲットが地球の裏側にある場合、待ち時間は100ms（往復200ms）であり、ローカルターゲットと通信する場合よりも非常に低速であろう。

【0160】

1つの実施形態において、2つの（又はそれ以上の）ローカル仮想ターゲットをミラーリングする場合、前述のように、コマンドの送信後に、本発明の実施形態によるスイッチは、イニシエータ（例えば、サーバ）から書込みデータを要求する前に、全てのターゲットからのR2T又はXFR\_RDYを受信するのを待つことになる。その後、書込みデータは、全てのターゲットにマルチキャストされる。しかしながら、低速リンク上でのミラーリングに関しては、長いネットワーク待ち時間を避けるために、スイッチは、リモートターゲットからR2T又はXFR\_RDYを受信するのを待たない。その代わりに、スイッチは、ローカルターゲットからR2T又はXFR\_RDYを受信すると、直ちに書込みデータをイニシエータから要求してローカルターゲットに書込む。リモート装置に接続するラインカードは、リモートターゲットからR2T又はXFR\_RDYを受信すると、ローカルターゲットからデータを読み出し、次に、データをリモートターゲットに書込む。

【0161】

詳細には、図25を参照すると、スイッチは、書込みコマンドをサーバから受信することになる（ステップ2502）。ローカルミラーリングの場合と同様に、入口側PPUは、コマンドをローカルターゲット及びリモートターゲットの両方の出口側ラインカードにマルチキャストすることになる（ステップ2504）。しかしながら、リモートターゲット宛のコマンドのフローIDは特別なフローIDなので、パケットは、他の環境で行われるように、PPUによって直接処理せ



ずに、出口側ラインカードCPUに導かれることになる。ローカルターゲット宛の packets は依然として PPU によって処理される。次に、コマンドは、それぞれの出口側ラインカードによって、各々のターゲットに、即ちローカルターゲット及びリモートターゲットに送られる（ステップ2506）。

#### 【0162】

ネットワーク待ち時間に起因して、R2T又はXFR\_RDYは、最初に、ローカルターゲットからスイッチによって受信されることになる（ステップ2508）。次に、R2T又はXFR\_RDYは、イニシエータ（サーバ）に返送されることになる（ステップ2510）。次に、イニシエータは、自身の書込みデータをスイッチに送信し、次に、データは、書込みのためにローカルターゲットに送られることになる（ステップ2512）。ローカルターゲットでの書込みが終了すると、ローカルターゲットは、タスクが完了したことを示す応答パケットを送信することになる（ステップ2514）。

#### 【0163】

最終的には、R2T又はXFR\_RDYは、ラインカードによってリモートターゲットから受信される（ステップ2516）。リモートターゲットに接続するラインカードのCPUは書込みコマンドを送信したので、リモートR2T又はXFR\_RDYは同様にラインカードCPUによって受信され、ラインカードCPUは、リモートターゲットへのコマンドを管理する点に留意されたい。リモートターゲットに関するラインカードCPUは、予め書込まれたデータを読取るために、受信したR2T又はXFR\_RDYをローカルターゲットに対する読み出しコマンドに変換する（ステップ2518）。ローカルターゲットから受信した読み出しデータは、リモートターゲットに関するラインカードのPPUによって受信される（ステップ2520）。次に、PPUは、読み出しデータを書込みデータとしてリモートターゲットへ送る（ステップ2522）。書込みが完了すると、リモートターゲットは、リモートターゲットに関するラインカードCPUによってどのパケットが受信されたかを示す応答パケットを送信することになる（ステップ2524）。ラインカードCPUは、読み出しコマンド及び書込みコマンドの両方に関する状態信号を受信する。

## 【0164】

ローカル書込みが完了する前にリモートターゲットのR2T又はXFR\_RDYを受信した場合、リモートラインカードは、1つの実施形態において、ローカルターゲットからのデータの読み出し処理を行う前に、ローカル書込みが完了するまで待機する。

## 【0165】

読み出し又は書込みのいずれかにエラーが発生した場合、ラインカードCPUは、エラーをSCCに報告する。エラー発生の場合、リモートターゲットは、ローカルターゲット及びラインカードに対して非同期になる。

## 【0166】

従って、ローカルターゲットに関して、書込みコマンドは、ローカルターゲットのラインカードのPPUで実行される。しかし、リモートターゲットに関して、書込みコマンドは、そのラインカードのPPUが読み出しデータを書込みデータとして送る以外は、リモートターゲットのラインカードのCPUによって管理される。

## 【0167】

スナップショット：

「スナップショット」とは、一般に、特定の時点まで仮想ターゲットをミラーリングし、その後、ミラーリングされたメンバを切断することにより、切断時点でミラーリングされたメンバのミラーデータをフリーズ(freeze)することである。換言すると、特定の時点でのデータの表面上の「スナップショット」が保持される。スナップショットが取得されると、ユーザは、リストアを必要とすることなく、(別の仮想ターゲットとしての)取り除かれたメンバにアクセスして、いつでも古い情報を検索することができる。従って、「スナップショット」を利用することで、本発明によるスイッチの一部のユーザは、従来のバックアップ作業及びリストア作業を行なう必要がなくなるであろう。更に、本発明によるスイッチを使用することによって、スナップショットを素早く行うことができ、所要時間は、仮想ターゲットをテープ媒体にコピーするために何時間ものバックアップウィンドウを必要とする(、及び通常はコピーされたデータへのアクセスも防

止する) 場合がある従来のバックアップに比較して、わずか数ミリ秒である。また、仮想ターゲットのスナップショットは、一定の時間間隔で行うことができる。更に、各々のスナップショットは、ミラーリングされた仮想ターゲットの異なるメンバとすることができ、複数のスナップショット(例えば、火曜日のスナップショット、水曜日のスナップショット等)の利用可能性を最適化するものである。

#### 【0168】

特に、図26を参照すると、本発明の1つの実施形態によるスナップショットサービスを行うために、スナップショット要求は、管理ステーションからスイッチによって受信される(ステップ2602)。SCCは、入口側ラインカードCPU(サーバに接続するラインカード)に、ミラーリングされたメンバを取り除くための変更を知らせる(ステップ2604)。また、SCCは、SCCデータベースの仮想ターゲットオブジェクトを更新する。ラインカードCPUは、もはや取り除かれたメンバを反映しないように仮想ターゲットの(PPU SRAM内にある)VTDに記憶されたフローIDを更新する(ステップ2606)。この変更により、受信した書込みデータは、取り除かれたメンバに対してマルチキャストされない。VTDが更新されると、CPUは、SCCに対する変更を了解し、SCCは、スナップショットが完了したことを示す応答信号を管理ステーションに返送する(ステップ2608)。

#### 【0169】

更に、任意のスナップショットを開始する前に、仮想ターゲットに対する未処理要求があってははいけない。従って、スナップショットが行われる場合、1つの実施形態において、仮想ターゲットに対する全ての未処理要求を休止するようにサーバに通知する必要がある。サーバの作動は、スナップショット後に再開される。

#### 【0170】

仮想ターゲットクローニング(複製) :

本発明によるスイッチは、ミラーリングされた仮想ターゲットへの新しいメンバの追加をサポートすることができ、本明細書ではクローニング(又は複製)と呼

び、オフラインで仮想ターゲットを取得することなく行なうことができる。一般に、新メンバは、SCCデータベースの仮想ターゲットオブジェクトを変更することによって追加され、ミラーリングされたターゲットの内容は、新メンバに複製されるが、仮想ターゲットに対する通常のアクセスは依然として有効である。仮想ターゲットのサイズにもよるが、複製を完了するにはある程度の時間を必要とするはずである。しかしながら、複製は、スイッチによって制御され、ユーザには見えず、一般にサーバによる仮想ターゲットへのアクセスを妨害しない。

#### 【0171】

詳細には、図27を参照すると、複製要求はSCCによって受信される（ステップ2702）。SCCは、クローニング継続中のフラグを仮想ターゲットオブジェクトに設定し（ステップ2704）、サーバに接続するラインカードのCPUに変更を知らせる（ステップ2706）。ラインカードCPUは、PPU SRAM内のVTDを更新し、仮想ターゲットのフローIDを変更して新メンバを追加する（ステップ2708）。フローIDの変更により、受信書込みデータはこの時点でマルチキャストされる。しかしながら、受信書込みはマルチキャストされるが、フローIDは新メンバに関する出口側ラインカードCPUへ書込みデータを導くように設定され、その結果、PPUのかわりにCPUが書込みデータを処理する。以下に詳細に説明するように、出口側ラインカードCPUは、複製が完了するまで新メンバに対するトラフィックを一時的に管理することになる。

#### 【0172】

新メンバに接続するラインカードのCPUは、新メンバにコピーされることになる仮想ターゲットの内容を指定する変更記述子を作成する（ステップ2710）。記述子は、オフセット及びブロックカウント（オフセット、ブロックカウント）を示す。例えば、10GBターゲットをコピーするための変更記述子は（0, 20, 000, 000）であり、1つの実施形態において、各々のブロックは512バイトであり、10GBターゲットは2000万個のブロックを有する点に留意されたい。ラインカードCPUは、変更記述子を使用して、1度に数ブロックのコピー機能を管理する。最初に、ラインカードCPUは、書込みコマンドを新メンバに送信する（ステップ2712）。R2T又はXFR\_\_RDYが

返送されると（ステップ2714）、ラインカードCPUは、旧メンバに対する読み出し要求を初期化するが、読み出しデータを新メンバのラインカードCPUに導くフローIDを指定する（ステップ2716）。何らかの読み出し又は書込みエラーが発生すると、コピーは異常終了してSCCに報告される。

#### 【0173】

変更記述子はブロック集合のコピー後に更新される（ステップ2718）。例えば、50個のブロックのコピー後に、前述の変更記述子は、最初の50個のブロックがこの時点では同期状態にないので（50, 19, 999, 950）になる。ブロック集合のコピー処理は、全ブロックがコピーされるまで続く（ステップ2720）。

#### 【0174】

仮想ターゲットが複数の領域で構成され、各々の領域が異なるラインカードを介してスイッチに結合されている場合、両方の領域に関する複製処理を同時に実行することができる。しかし、両方の領域が同じラインカードを介してスイッチに結合されている場合、複製プロセスは、順次実行する必要がある。即ち、第1の領域の複製が完了するまで、第2の領域は複製できない。

#### 【0175】

一時的に、複製処理の間、仮想ターゲットへの書込み要求は、サーバから受信することができるとともに全てのミラーリングされたメンバに書込む必要があり、受信処理において仮想ターゲットの全データであるメンバを含む。この場合、書込み要求がマルチキャストされると、その書込み要求はミラーリングされたターゲットの旧メンバと同様に、それぞれのラインカード上のPPUによって処理されるのではなく、新メンバのラインカードのCPUによって受信される（ステップ2722）。ラインカードCPUは、書込み位置を変更記述子のオフセットと照らし合わせることによって、書込みが未コピーのブロックのいずれかに対するものであるか否かを判定する（ステップ2724）。書込みがコピー済みのデータブロックに対するものである場合、書込みコマンドは、単純にPPUに送られる（ステップ2726）。しかしながら、書込みが未コピーのデータブロックに対するものである場合、新メンバに対する書込みは破棄され（ステップ272

8)、そしてイニシエータに対してタスク完了の応答信号を送信する。それでもなお、新しいデータは、最終的には続行している複製処理中に旧メンバから新メンバへコピーされることになる。この処理は、完了するまで複製を実行し続ける(ステップ2720)。

#### 【0176】

別の方法において、複製処理中に仮想ターゲットに対する書込み要求が受信されると、仮想ターゲットに対して行われた変更は、ラインカードCPUによって追跡することができる。複製が完了すると、その変更及び追跡部分を更新できる。

#### 【0177】

複製処理が完了すると、ラインカードCPUは、SCCに通知する(ステップ2730)。SCCは、クローニング継続中のフラグを解除するために仮想ターゲットオブジェクトを更新する(ステップ2732)。イニシエータに接続する入口側ラインカード上ではフローIDが更新されるので、書込みコマンドは、新メンバのラインカードCPUに導かれるのではなく、通常どおりPPUへ進む(ステップ2734)。

#### 【0178】

##### 第三者コピー：

第三者機能は、オフライン仮想ターゲット(アクセスされていないもの)を書込み可能CD又はテープドライブ等のアーカイブ装置へ又はアーカイブ装置からコピーする。このコピーは、コピーが完了するまでサーバは関与せず、むしろスイッチによって実行されるので「第三者コピー」と呼ばれる。多くの実施形態において、このような第三者コピーは、予め取得した仮想ターゲットのスナップショットから行われることになる。従来システムの大部分においては、このようなコピーを行うために、ターゲット装置はスマートテープ装置等の「スマート」な装置である必要があり、このことは、装置が全体的にコピー処理に積極的に関与し、且つ少なくとも部分的にコピー処理を制御することを意味する。対照的に、本システムの第三者コピーサービスは、記憶装置スイッチ外部の処理能力によるものではない。

#### 【0179】

図28を参照すると、スイッチは、コピー要求を管理ステーションから受信することになる（ステップ2802）。SCCは、仮想ターゲットへの書込みに関する未処理接続が確実に存在しないようにする（ステップ2804）。コピー時に、仮想ターゲットは、1つの実施形態において、読み出しにのみ利用可能である。次に、SCCは、SCCデータベースの仮想ターゲットオブジェクトにコピー継続中のフラグを設定して、ターゲットへの書込みに関する他の接続が確実に存在しないようにする。次に、SCCは、コピー宛先装置に接続されたラインカードのCPUにコピーを実行するよう指示する（ステップ2808）。

#### 【0180】

各々の仮想ターゲットは複数の領域から構成することができ、各々の領域は異なる物理的装置上にあってもよい。従って、宛先ラインカードのCPUは、各々の領域からデータを取得する必要がある。これを行うために、宛先ラインカードのCPUは、各々の領域の各々のラインカードに領域記述子を送信する（ステップ2810）。領域記述子は、領域及び宛先ラインカード（宛先コピー用）を指定する。次に、それぞれの領域に関するラインカードの各々のCPUは、それぞれのPPU（例えば、VTD及びCAM）をセットアップして、PPUが読み出し要求を処理できるようにする（ステップ2812）。

#### 【0181】

領域ラインカードがセットアップ状態になると、次に、宛先ラインカードCPUは、書込みコマンドを宛先装置に送信する（ステップ2814）。R2T又はXFR\_RDYを宛先ラインカードによって宛先装置から受信すると（ステップ2816）、宛先ラインカードは、それぞれの領域ラインカードを経由して読み出しコマンドを領域の1つに送る（ステップ2818）。読み出しデータは、宛先ラインカードに直接送信され、宛先ラインカードPPUによって書込みデータとして処理され（ステップ2820）、書込みデータは、宛先装置に書込まれる。この処理は、領域全体がコピーされるまで繰り返される。何らかのエラーが発生するとコピーは終了する。次に、全ての領域がコピーされていない場合（ステップ2822）、この処理はステップ2814に戻り、次の領域のコピーが実行される。全ての領域がコピーされた場合（ステップ2822）、宛先ラインカード

のCPUは、コピー完了をSCCに報告する(ステップ2824)。エラー完了の場合、SCCは、コピーを終了する。しかし、コピーがエラーなしで完了した場合、SCCは、SCCデータベースの仮想ターゲットオブジェクトのコピー継続中のフラグをリセットし(ステップ2826)、管理ステーションに完了状態を報告する(ステップ2828)。ソース仮想ターゲットは、この時点で再び書込みができるようになる。

#### 【0182】

##### 定期的スナップショット及びバックアップ:

本発明の実施形態によるスイッチは、仮想ターゲットの定期的スナップショット及びバックアップを行うことができる。このようなバックアップ機能は、一般に3つのステップを含む。

1. 仮想ターゲットをスナップショットする。
2. スナップショットから仮想ターゲットを第三者コピーする。
3. 仮想ターゲットへスナップショットを送るメンバをミラーリングされたメンバとして再加入させて、最新のミラーリングされたデータの全てをそのメンバに持ち込む。

#### 【0183】

第3のステップは、(前述の)複製によって、又は、スナップショットが取得された時間からメンバが再加入されるまで、仮想ターゲットの更新データを追跡する他の方法によって実行することができる。例えば、仮想ターゲットに対して行われた全ての変更記録を保持することができ、そして次に、ミラーリングされたメンバは、ミラーリングされたメンバとして仮想ターゲットを再加入させた時点で、単純にこれらの変更内容で更新される。

#### 【0184】

ユーザが多数の記憶空間を有する場合、ユーザは各々のスナップショット仮想ターゲットにアクセスできるはずなので、第2のステップ及び第3のステップは必要ない場合もある。従って、このことはスナップショットターゲットを割り当て、且つネーミングを行うという問題に過ぎない。例えば、今週は就業日毎に、過去半年は月別に、その後は、四半期別に仮想ターゲットをバックアップするこ



とになっている場合、有限のスナップショットターゲット集合のみを割り当てる必要があり、以下のように命名できる。

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . m a s t e r

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . m o n d a y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . t u e s d a y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . w e d n e s d a y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . t h u r d a y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . f r i d a y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . f e b r u a r y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . m a r c h

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . a p r i l

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . m a y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . j u n e

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . j u l y

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v  
e r . b a c k u p . 2 0 0 0 q 3

i q n . c o m . m a r a n t i n e t w o r k s . c o m p a n y . s e r v

er. backup. 2000q4

iqn. com. marantinenetworks. company. server.

er. backup. 2001q1

iqn. com. marantinenetworks. company. server.

er. backup. 2001q2

#### 【0185】

スイッチは、スナップショットターゲットを割り当て、且つ既知のポリシーに基づいて定期的な作動のスケジュールを組む。また、スイッチは、ターゲットのネーミング（命名）及びリネーミング（改名）を管理する。例えば、バックアップ2001q3については、スイッチは、backup. 2000q3のターゲットを再使用して、backup. 2001. q3のためにネーミングを行うことになる。

#### 【0186】

##### リストア：

種々の理由から、多くの業界では、アーカイブ媒体（例えば、テープ又はCD等の一般的に取り外し可能な媒体又は携帯式媒体）にデータのバックアップコピーを保存する必要がある。スイッチは、第三者コピー機能を利用して、バックアップ又はスナップショットターゲットをアーカイブ媒体に移すことができる。スイッチは、アーカイブ媒体をデータベース上で追跡する。アーカイブ媒体のコピーが行われる度に、SCCは、全ての宛先領域を判定するために仮想ターゲットオブジェクトを取り込み、そして、記録は、媒体を追跡するために管理ステーションでデータベースに入力される。管理ステーションを使用して、ユーザは、テープ又はCD集合のアーカイブ媒体のリストを閲覧し、リストアのために1つを選択することができる。

#### 【0187】

リストア操作自体は、スイッチによってスケジュールされる別の第三者コピー機能である。しかしながら、この操作は、誰かが媒体をテープ又はCDドライブに入れる必要があるのでユーザの介入を伴う。しかしながら、本明細書で説明した他の記憶装置の場合と同様に、ソースターゲット装置のCPUは、リストア操作の作業を制御するが、複数の宛先SPUは1つずつ必要とされる。

## 【0188】

本発明の1つの実施形態によるスイッチは、リストア処理の3つの異なる優先順位、即ち、緊急、重要、及び通常をサポートする。緊急リストアは、システム上の現在のトラフィック状況に無関係に直ちに開始される。重要リストアは、システムが混雑している場合には開始されないが、数時間以内に開始される。通常リストアは、システムのトラフィック混雑に応じて24時間以内に完了される。

## 【0189】

結論：

従って、本発明の実施形態に基づいて、パケットの分類、パケット上での仮想化機能の実行、パケットの任意の所要プロトコル変換の実行を含む、データパケットのワイヤ速度での処理を可能にする記憶装置スイッチが開示される。従来の方法と比較すると、開示されたアーキテクチャによって、パケットを処理するために必要な時間を最小にすることができる。このようなワイヤ速度での処理は、ある意味では、スイッチの処理能力を全てのラインカードに分散して、全体的にバッファリングの必要性を回避することによって達成される。このような分散された処理能力は、高帯域幅を有するだけでなく、拡張が容易なシステムを可能にする。更に、自身のラインカードを使用するこのようなスイッチは、サーバ不要の記憶サービス、即ち、スイッチの外部にはその運用の制御に必要な実体物が全くないサービスを行うこともできる。

## 【0190】

以上説明した特定の実施形態は、本発明の原理を例示するものに過ぎず、当業者であれば、本発明の技術範囲及び技術思想から逸脱することなく種々の変更を行うことができることを理解されたい。従って、本発明の技術範囲は、特許請求の範囲によってのみ限定される。

## 【図面の簡単な説明】

## 【0191】

【図1】従来のシステムによるSANの概略機能ブロック図である。

【図2】従来の方法によるプロトコル間のインタフェース処理に使用される装置の概略機能ブロック図である。

【図 3】本発明の実施形態による記憶装置スイッチを使用する SAN システムの概略機能ブロック図である。

【図 4】本発明の実施形態による記憶装置スイッチを使用するシステムの別の実施形態の概略機能ブロック図である。

【図 5】本発明の実施形態による記憶装置スイッチを使用するシステムの更に別の実施形態の概略機能ブロック図である。

【図 6】本発明の実施形態による記憶装置スイッチの概略機能ブロック図である。

【図 7】本発明の実施形態による記憶装置スイッチに使用されるラインカードの概略機能ブロック図である。

【図 7 a】本発明の実施形態による記憶装置スイッチで使用される仮想ターゲット記述子の概略ブロック図である。

【図 8 a】従来公知の iSCSI PDU の概略ブロック図である。

【図 8 b】従来公知の iSCSI PDU の概略ブロック図である。

【図 8 c】従来公知の iSCSI PDU の概略ブロック図である。

【図 8 d】従来公知の iSCSI PDU の概略ブロック図である。

【図 8 e】従来公知の iSCSI PDU の概略ブロック図である。

【図 8 f】従来公知のファイバ・チャンネルプロトコル (FCP) フレーム及びペイロードの概略ブロック図である。

【図 8 g】従来公知のファイバ・チャンネルプロトコル (FCP) フレーム及びペイロードの概略ブロック図である。

【図 8 h】従来公知のファイバ・チャンネルプロトコル (FCP) フレーム及びペイロードの概略ブロック図である。

【図 8 i】従来公知のファイバ・チャンネルプロトコル (FCP) フレーム及びペイロードの概略ブロック図である。

【図 9 a】本発明の実施形態による、PACE において行われる、入口側の iSCSI パケット分類処理を示す流れ図である。

【図 9 b】本発明の実施形態による、PACE において行われる、出口側の iSCSI パケット分類処理を示す流れ図である。

【図10a】本発明による記憶装置スイッチに入るときのTCPパケット及びTCPパケットのブロック図であり、パケットが記憶装置スイッチ内での使用に適するように変更される方法を示す。

【図10b】本発明による記憶装置スイッチに入るときのTCPパケット及びTCPパケットのブロック図であり、パケットが記憶装置スイッチ内での使用に適するように変更される方法を示す。

【図11】本発明の実施形態による記憶装置スイッチで使用されるローカルヘッダの概略ブロック図である。

【図12a】本発明の実施形態による、PACEにおいて行われる、入口側のFCPフレーム分類処理を示す流れ図である。

【図12b】本発明の実施形態による、PACEにおいて行われる、出口側のFCPフレーム分類処理を示す流れ図である。

【図13a】本発明の実施形態による、PPUにおいて行われる、入口側の分類処理を示す流れ図である。

【図13b】本発明の実施形態による、PPUにおいて行われる、出口側の分類処理を示す流れ図である。

【図14】本発明の実施形態による、コマンドパケット又はフレームに関する入口側の仮想化処理を示す流れ図である。

【図14a】仮想化処理時のローカルヘッダ及びタスク制御ブロック（ITCB及びETCB）のブロック図であり、入口側での（イニシエータサーバ／ポートからの）コマンドパケットのヘッダ及びITCBを示す。

【図15】本発明の実施形態による、コマンドパケット又はフレームに関する出口側の仮想化処理を示す流れ図である。

【図15a】仮想化処理時のローカルヘッダ及びタスク制御ブロック（ITCB及びETCB）のブロック図であり、出口側での（ファブリック／トラフィックマネージャからの）コマンドパケットのヘッダ及びETCBを示す。

【図16】本発明の実施形態による、R2T/XFR\_RDYパケット又はフレームに関する入口側での仮想化処理を示す流れ図である。

【図16a】仮想化処理時のローカルヘッダ及びタスク制御ブロック（ITCB

及びETCB)のブロック図であり、入口側での(ターゲット記憶装置/ポートからの)R2T/XFR\_RDYパケットに関するヘッダ及びETCBを示す。

【図17】本発明の実施形態による、R2T/XFR\_RDYパケット又はフレームに関する出口側での仮想化処理を示す流れ図である。

【図17a】仮想化処理中のローカルヘッダ及びタスク制御ブロック(ITCB及びETCB)のブロック図であり、出口側での(ファブリック/トラフィックマネージャからの)R2T/XFR\_RDYパケットに関するヘッダ及びITCBを示す。

【図18】本発明の実施形態による、書込みデータパケット又はフレームに関する入口側での仮想化処理を示す流れ図である。

【図18a】仮想化処理時のローカルヘッダ及びタスク制御ブロック(ITCB及びETCB)のブロック図であり、入口側での(イニシエータサーバ/ポートからの)書込みデータパケットに関するヘッダ及びITCBを示す。

【図19】本発明の実施形態による、書込みデータパケット又はフレームに関する出口側での仮想化処理を示す流れ図である。

【図19a】仮想化処理時のローカルヘッダ及びタスク制御ブロック(ITCB及びETCB)のブロック図であり、出口側での(ファブリック/トラフィックマネージャからの)書込みデータパケットに関するヘッダ及びETCBを示す。

【図20】本発明の実施形態による、読み出しデータパケットに関する入口側での仮想化処理を示す流れ図である。

【図20a】仮想化処理時のローカルヘッダ及びタスク制御ブロック(ITCB及びETCB)のブロック図であり、入口側での(ターゲット記憶装置/ポートからの)書込みデータパケットに関するヘッダ及びETCBを示す。

【図21】本発明の実施形態による、読み出しデータパケットに関する出口側での仮想化処理を示す流れ図である。

【図21a】仮想化処理時のローカルヘッダ及びタスク制御ブロック(ITCB及びETCB)のブロック図であり、出口側での(ファブリック/トラフィックマネージャからの)書込みデータパケットに関するヘッダ及びITCBを示す。

【図22】本発明の実施形態による、応答パケット又はフレームに関する入口側

での仮想化処理を示す流れ図である。

【図22a】仮想化処理時のローカルヘッダ及びタスク制御ブロック（ITCB及びETCB）のブロック図であり、入口側での（ターゲット記憶装置／ポートからの）応答パケットに関するヘッダ及びETCBを示す。

【図23】本発明の実施形態による、応答パケット又はフレームに関する出口側での仮想化処理を示す流れ図である。

【図23a】仮想化処理中のローカルヘッダ及びタスク制御ブロック（ITCB及びETCB）のブロック図であり、出口側での（ファブリック／トラフィックマネージャからの）応答パケットに関するヘッダ及びITCBを示す。

【図24】本発明の実施形態による記憶サービスを実行するために行われるステップを示す流れ図である。

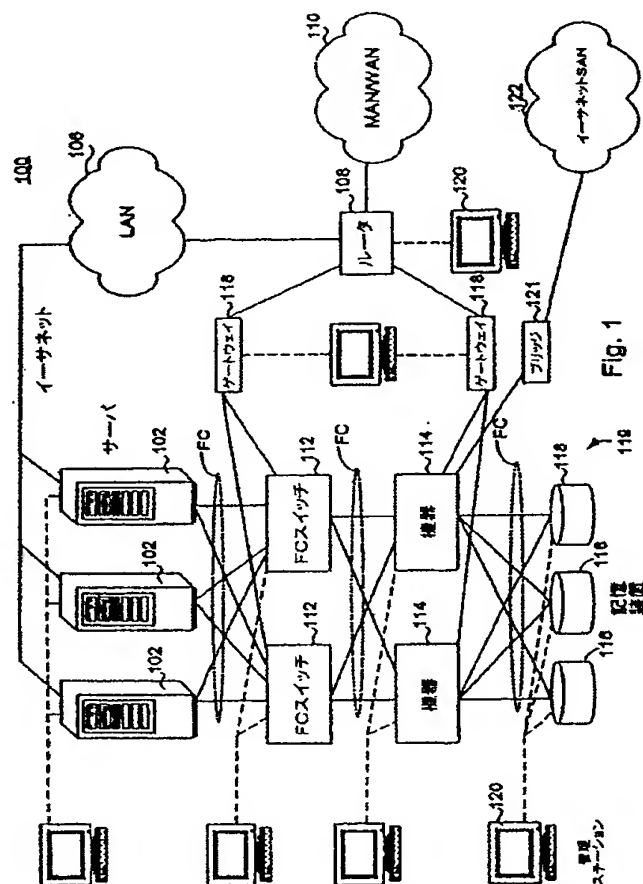
【図25】本発明の実施形態による低速リンク上のミラーリング記憶サービスのために行われる概略的ステップを示す流れ図である。

【図26】本発明の実施形態によるスナップショット記憶サービスのために行われるステップを示す流れ図である。

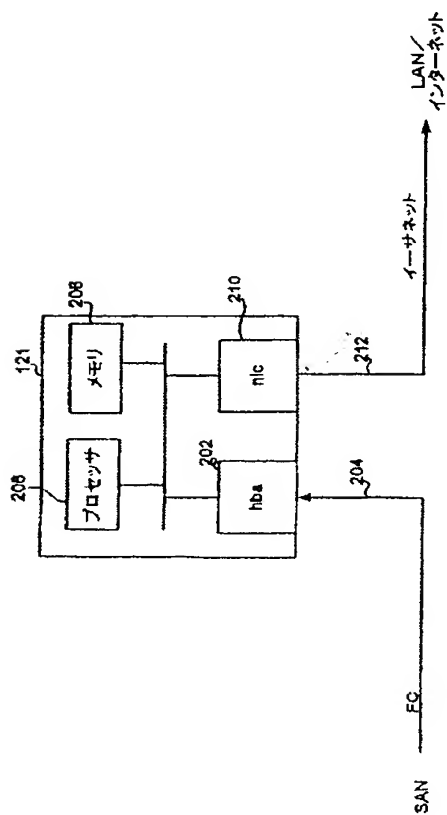
【図27】本発明の実施形態によるクローニング記憶サービスのために行われるステップを示す流れ図である。

【図28】本発明の実施形態による第三者コピー記憶サービスのために行われるステップを示す流れ図である。

【図1】



【図2】





【図3】

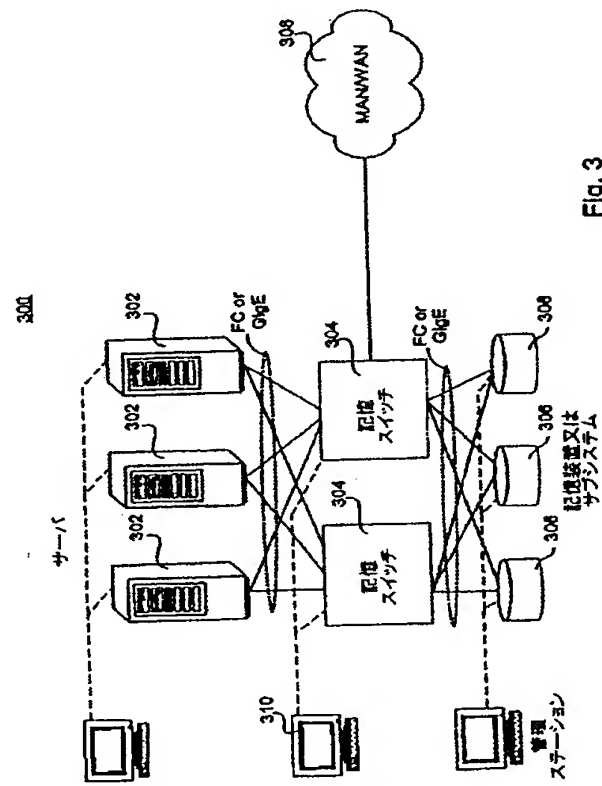


Fig. 3

【図4】

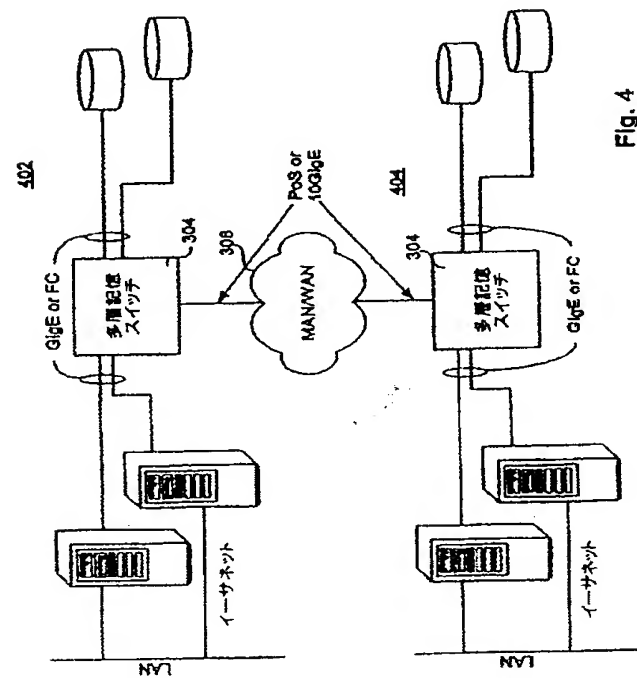


Fig. 4

【図5】

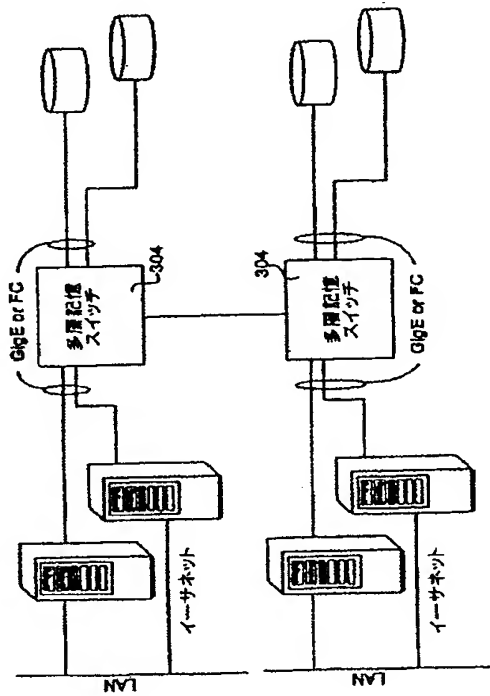


Fig. 5

【図6】

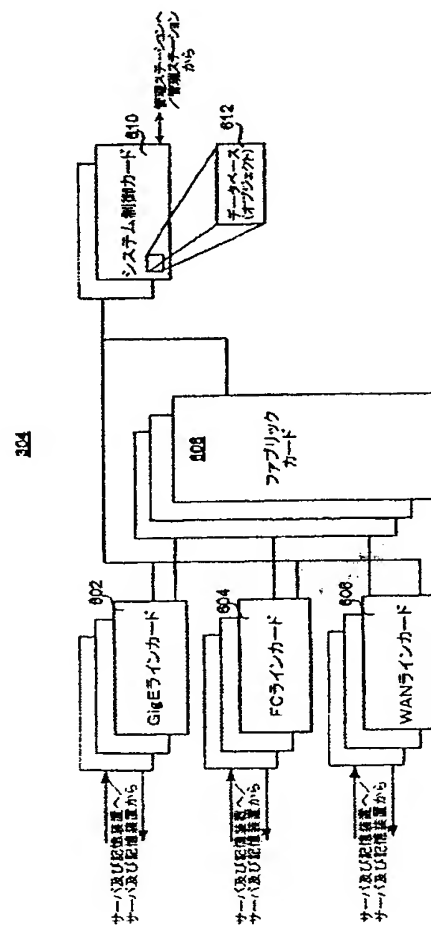
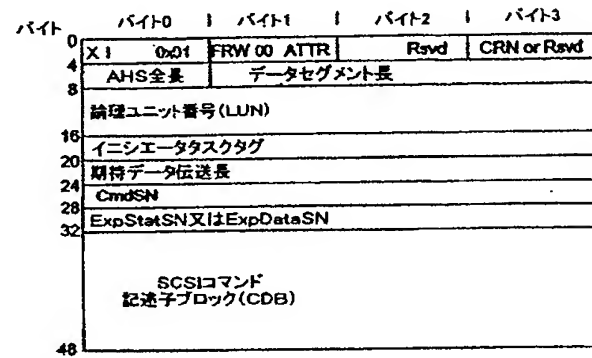


Fig. 6



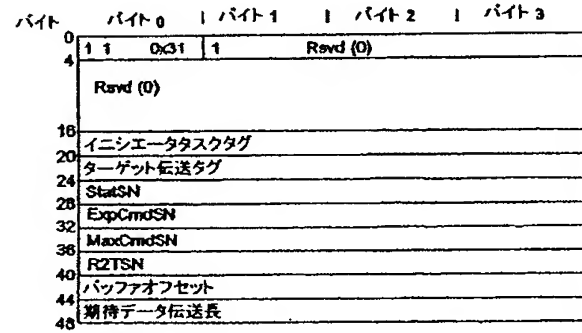
【図8a】



iSCSIコマンドPDU

Fig. 8a

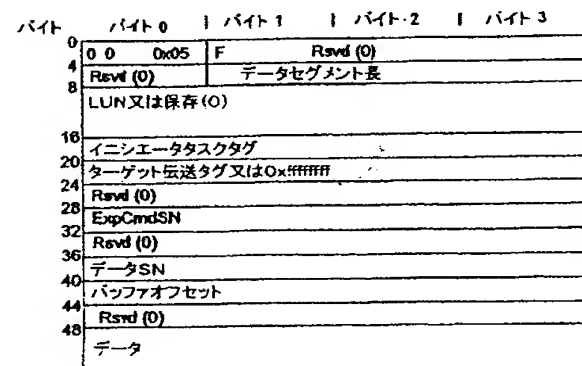
【図8b】



iSCSI R2T FDU

Fig. 8b

【図8c】



iSCSI書き込みデータPDU

Fig. 8c

【図8d】

バイト0	バイト1	バイト2	バイト3
1 1 0x25 F OUS Rsvd(0) 状態又はRsvd			
Rsvd(0)	データセグメント長		
Rsvd(0)			
イニシエータタスクタグ			
Rsvd(0)			
StatSN又はRsvd(0)			
ExpCmdSN			
MaxCmdSN			
データSN			
バッファオフセット			
残余カウンタ			
データ			

iSCSI読み取りデータPDU

Fig. 8d

【図8e】

バイト0	バイト1	バイト2	バイト3
1 1 0x21 1 rsv 0 u 0 # 0 状態 応答			
Rsvd(0)	データセグメント長		
Rsvd(0)			
イニシエータタスクタグ			
基本残余カウンタ			
StatSN			
ExpCmdSN			
MaxCmdSN			
ExpDataSN又はRsvd(0)			
ExpR2TN又はRsvd(0)			
Bidi-競出し残余カウンタ			
センスデータ及び応答データ(オプション)			

iSCSI応答PDU

Fig. 8e

【図8f】

ビット ワード	31-24	23-16	15-08	07-00
0	R_CTL		D_ID	
1	rsvd		S_ID	
2	TYPE		F_CTL	
3	SEQ_ID	DF_CTL	SEQ_CNT	
4		OX_ID	RX_ID	
5		RLTV_OFF		

FCフレームヘッダ

Fig. 8f

【図8g】

フィールド名	内容	サイズ
FCP_LUN	論理ユニット番号	8バイト
FCP_CNTL	制御フィールド	4バイト
FCP_CDB	SCSIコマンド記述子ブロック	16バイト
FCP_DL	データ長	4バイト

FCP\_CMNDペイロード

Fig. 8g

【図8h】

フィールド名	内容	サイズ
DATA_RO	連続のFCP DATA IUの 第1のバイトの相対オフセット	4バイト
BURST_LEN	連続のFCP DATA IUの長さ	4バイト
rsvd		4バイト

FCP\_XFR\_RDYペイロード

Fig. 8h

【図8i】

フィールド名	内容	サイズ
rsvd		4バイト
rsvd		4バイト
FCP_STATUS	フィールド有効性及びSCSI状態	4バイト
FCP_RESID	残余カウンタ	4バイト
FCP_SNS_LEN	FCP_SNS_INFOフィールド長	4バイト
FCP_RSP_LEN	FCP_RSP_INFOフィールド長	4バイト
FCP_RSP_INFO	FCP応答情報	mバイト
FCP_SNS_INFO	FCPセンス情報	nバイト

FCP\_RSPペイロード

Fig. 8i

【図9a】

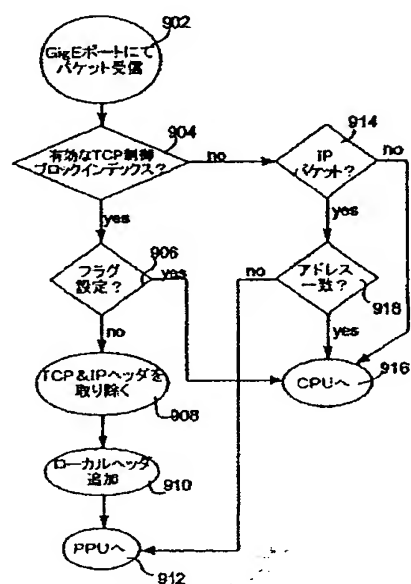
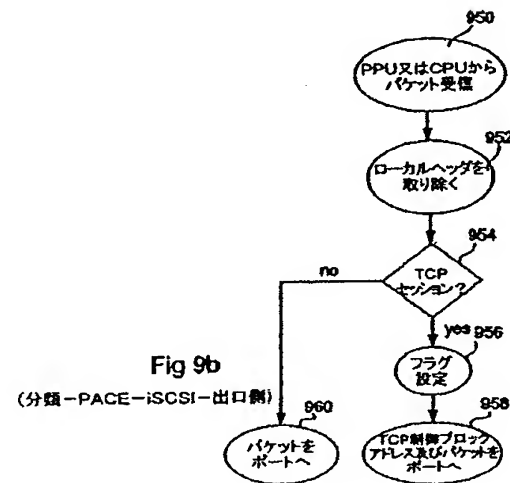


Fig 9a

(分類-PACE-iSCSI-入口側)

【図9b】



【図10a】

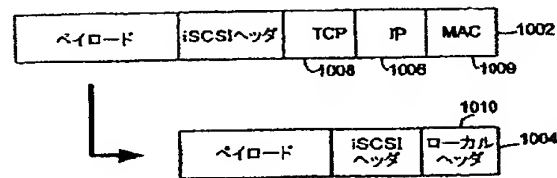


Fig. 10a

【図10b】

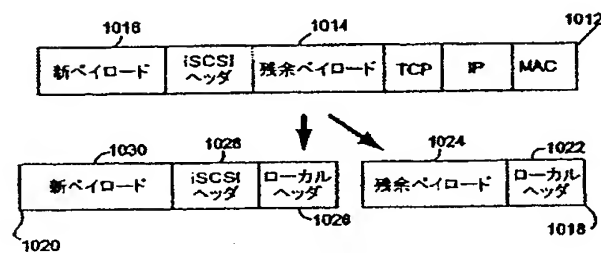


Fig. 10b

【図11】

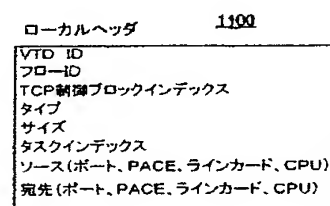
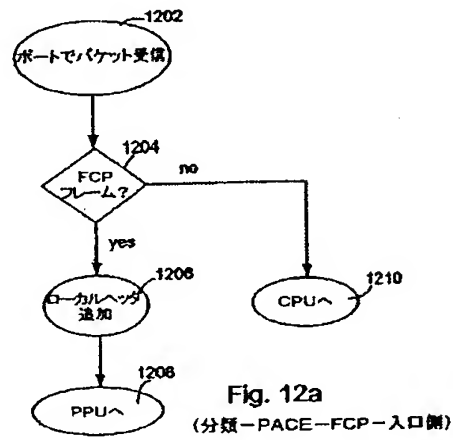
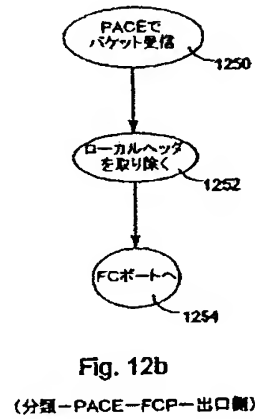


Fig. 11

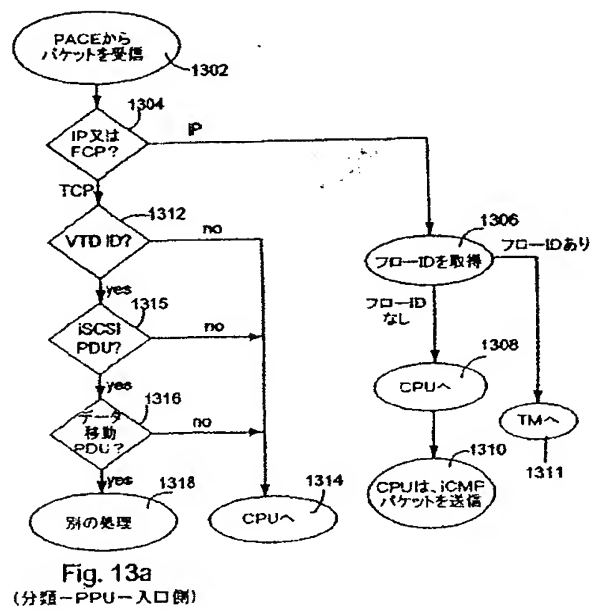
【図12a】



【図12b】

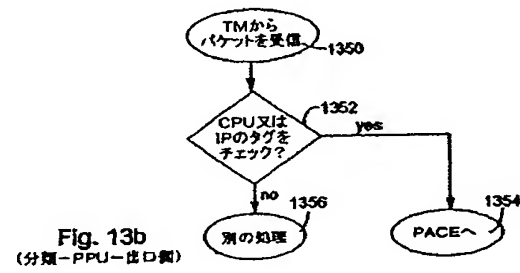


【図13a】





【図13b】



【図14】

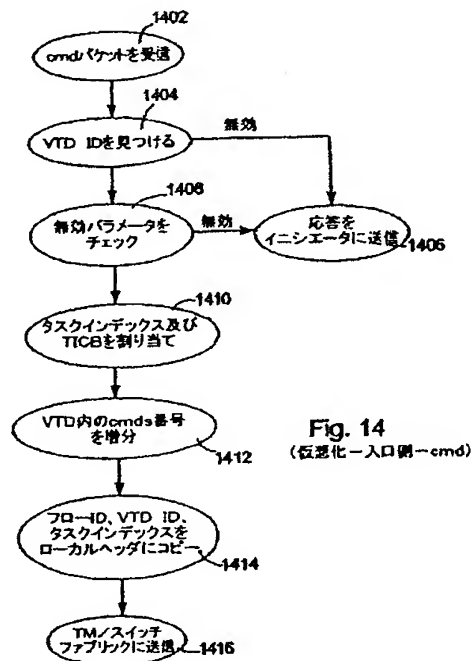


Fig. 15a  
出口機-Cond

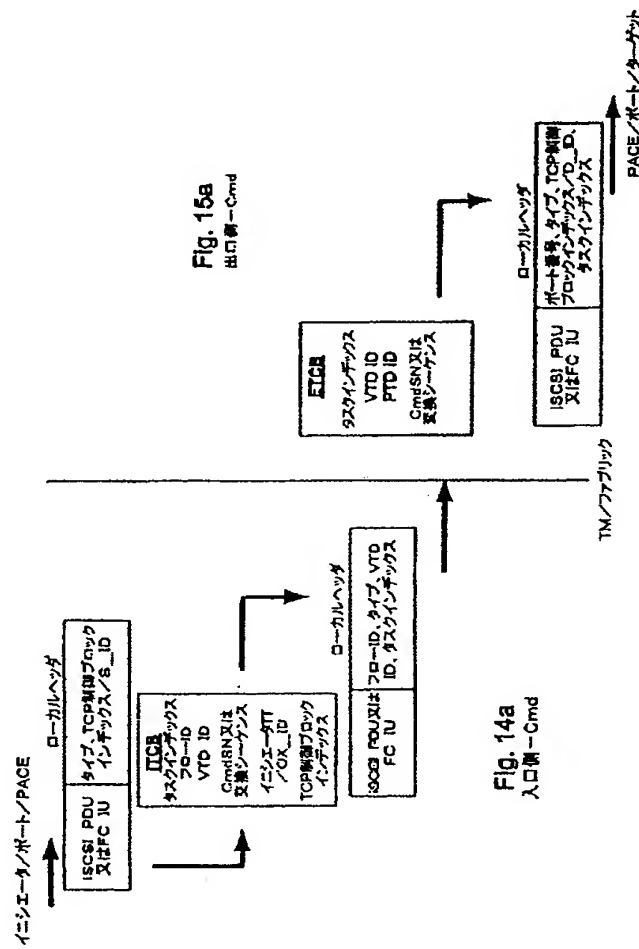
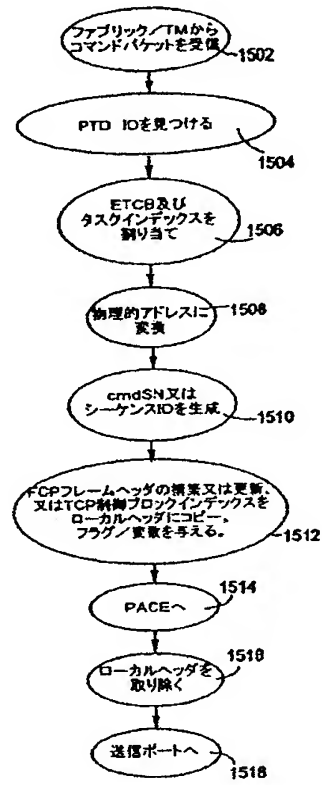
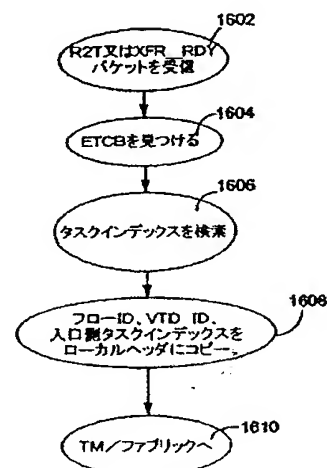


Fig. 14a  
入口側-Cmd

【図15】

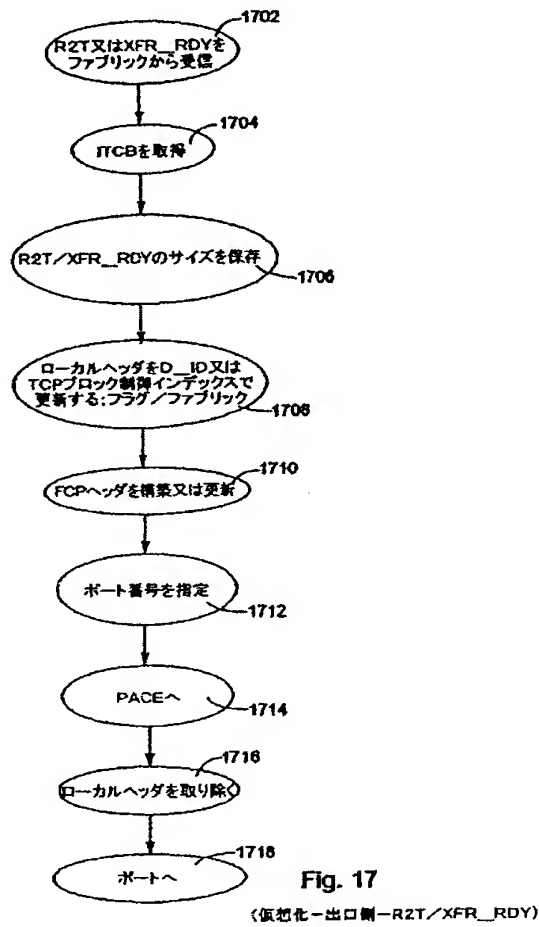
Fig. 15  
(仮造化-出口側-cmd)

【図16】

Fig. 16  
(仮造化-入口側-R2T/XFR\_RDY)



【図17】



【図18】

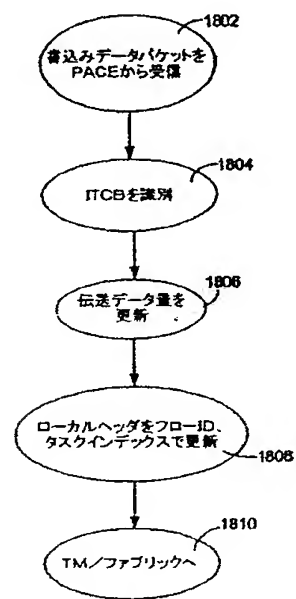
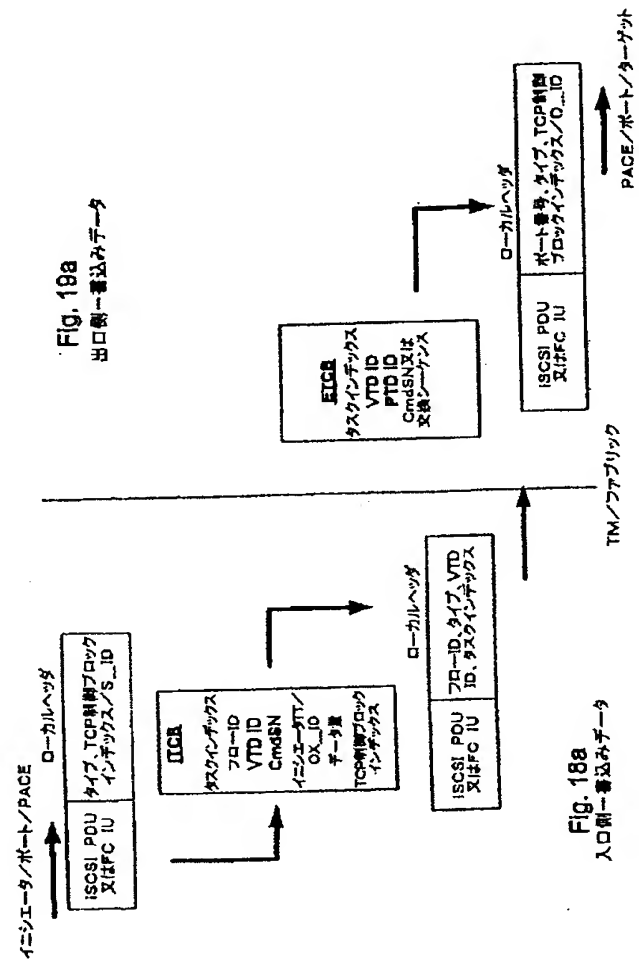


Fig. 18  
(仮想化-入口側-蓄込みデータバケット)

【図18a-19a】



【図19】

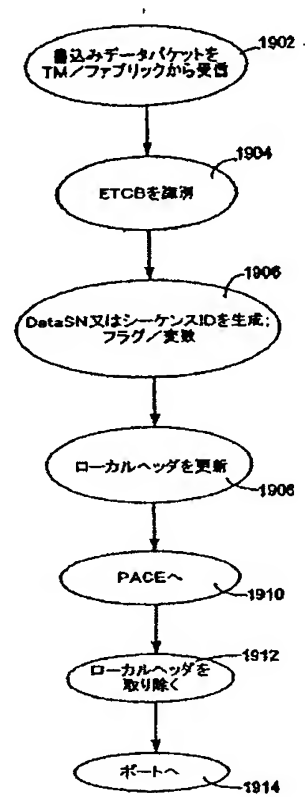


Fig. 19  
(仮想化-出口側-書き込みデータバケット)

【図20】

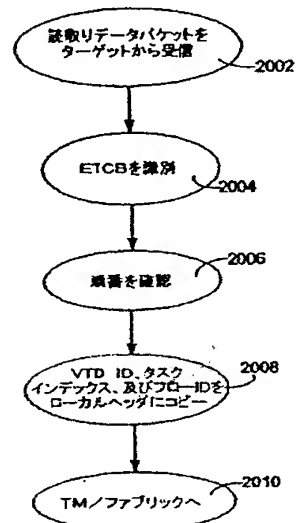
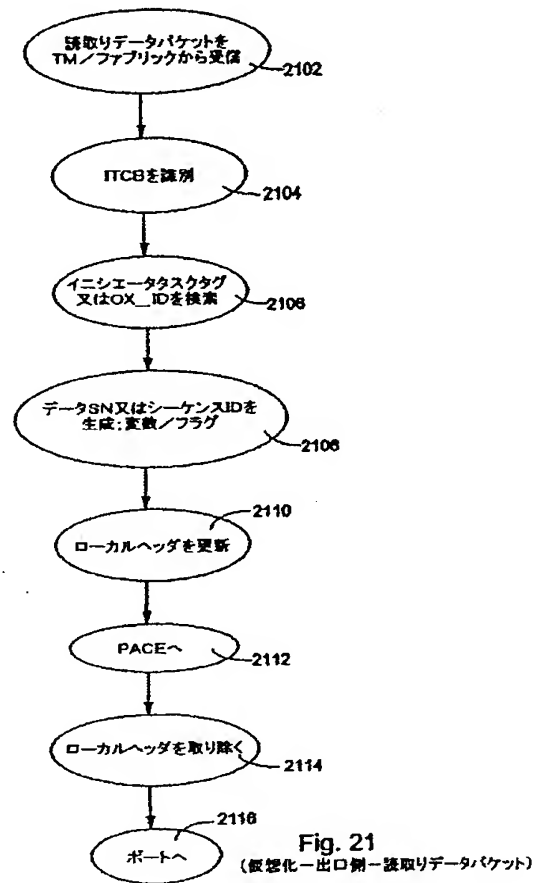


Fig. 20  
(仮想化-入口側-読み取りデータバケット)

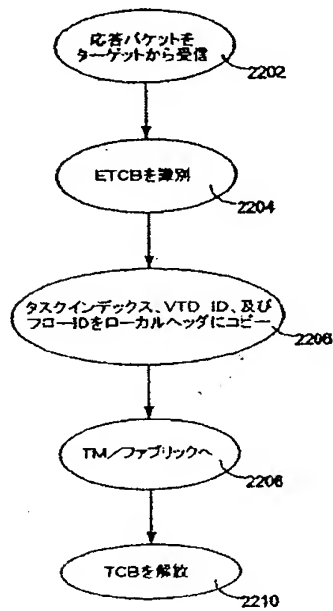




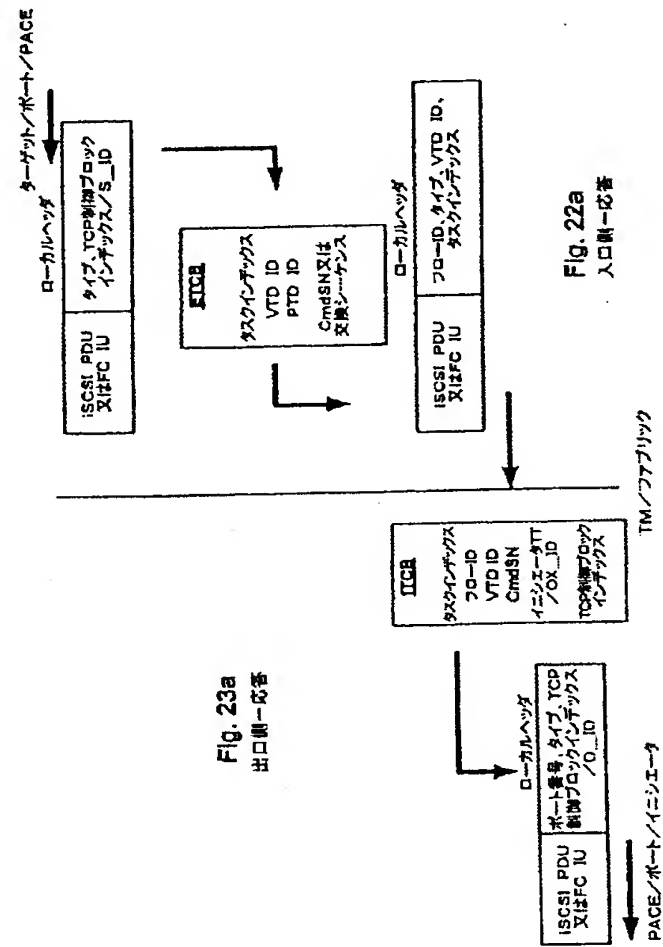
【図21】



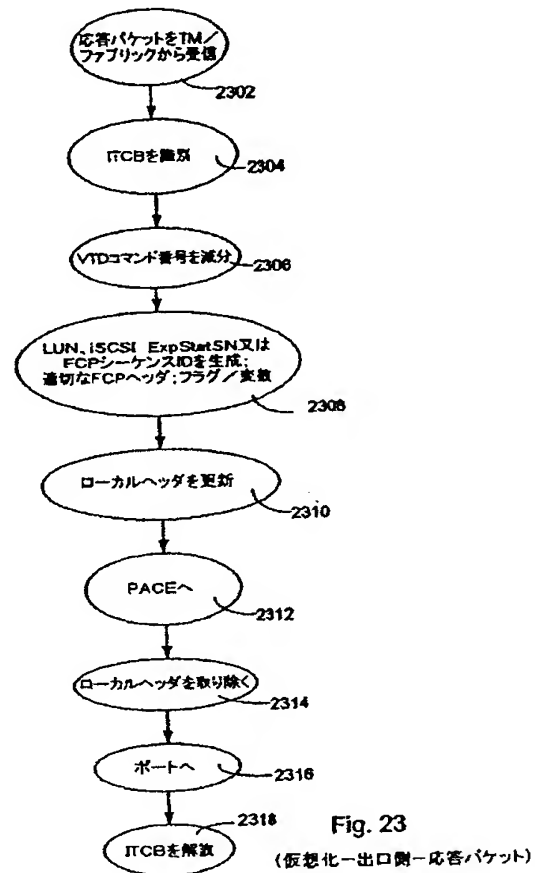
【図22】

Fig. 22  
(仮想化—入口側—応答パケット)

【図 2 2 a - 2 3 a】



【図23】



【図24】

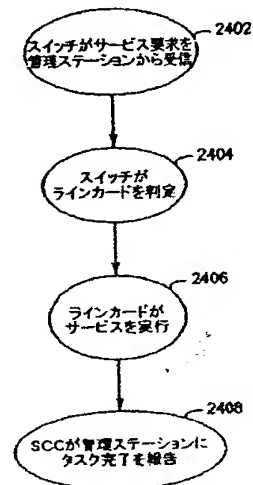


Fig.24

【図25】

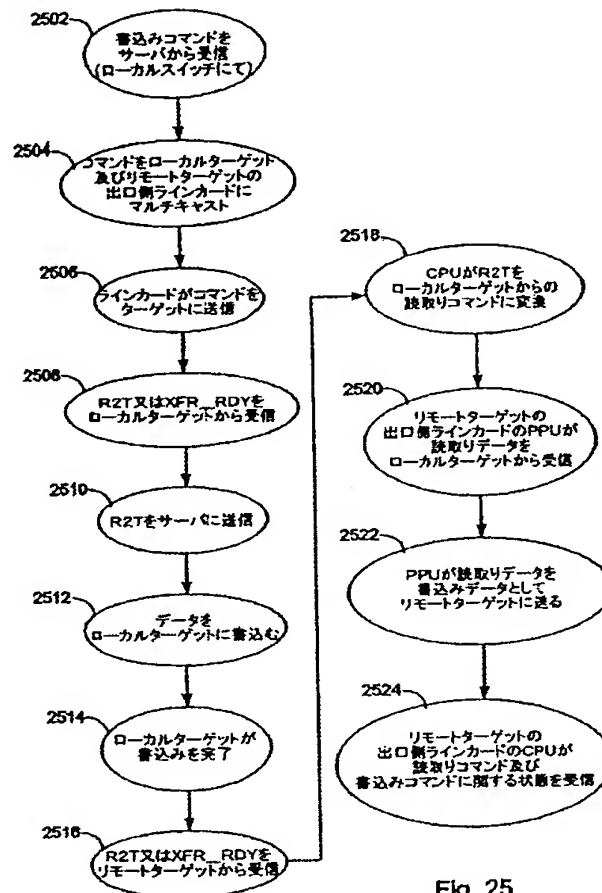


Fig. 25

【図26】

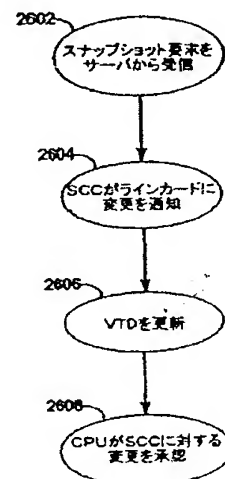
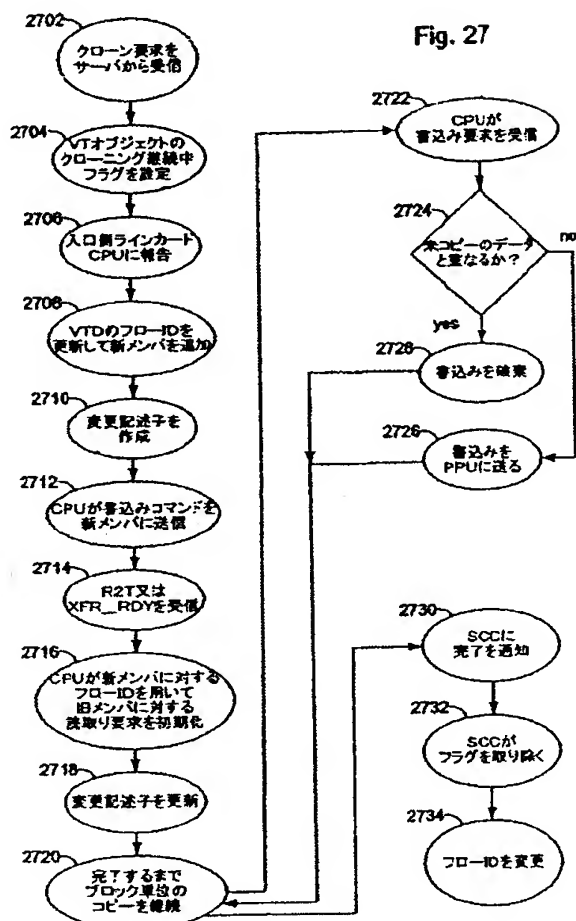


Fig. 26

【図27】



【図28】

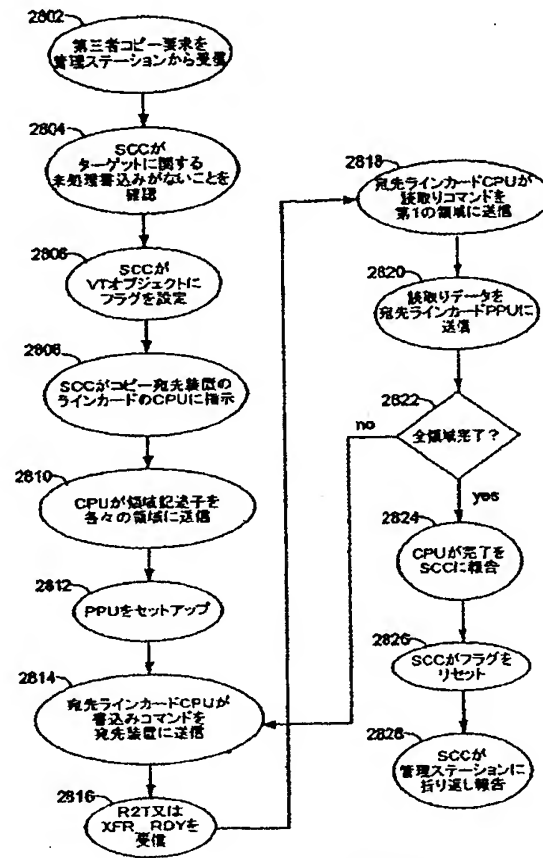
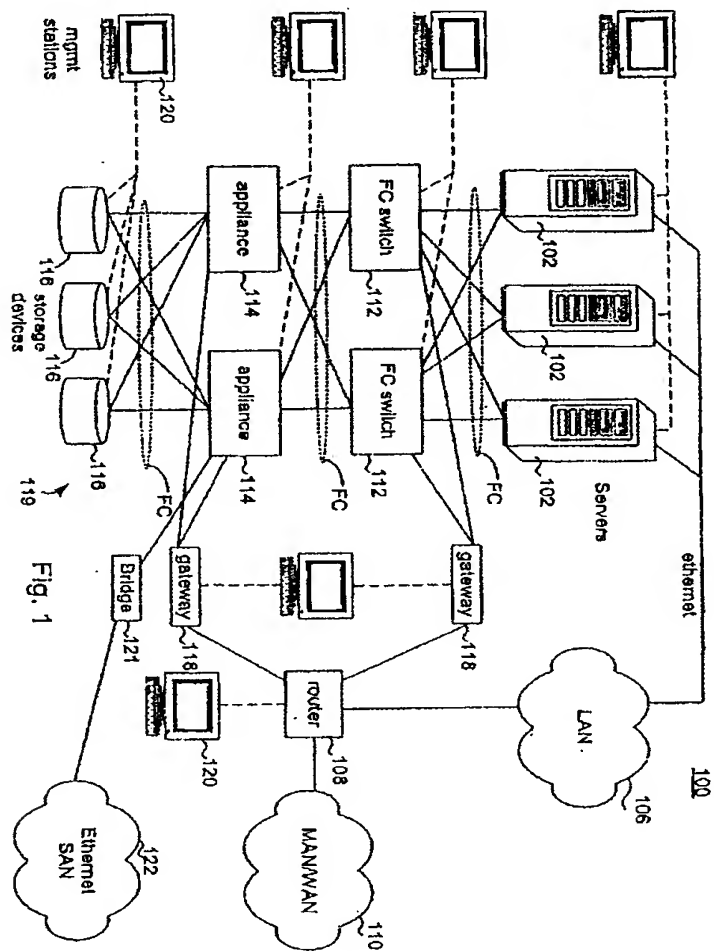


Fig. 28

WO 03/027877

PCT/US02/34912

1/38



WO 03/027877

PCT/US02/30912

2/38

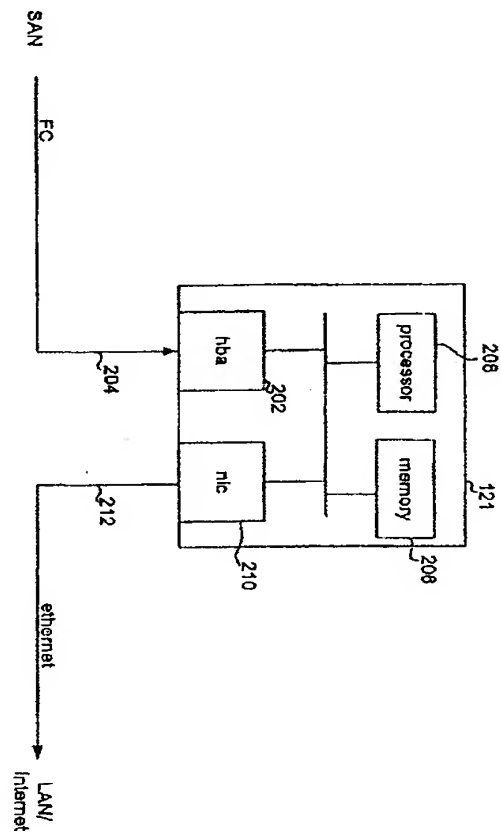


Fig. 2



WO 03/027877

PCT/US02/36912

3/38

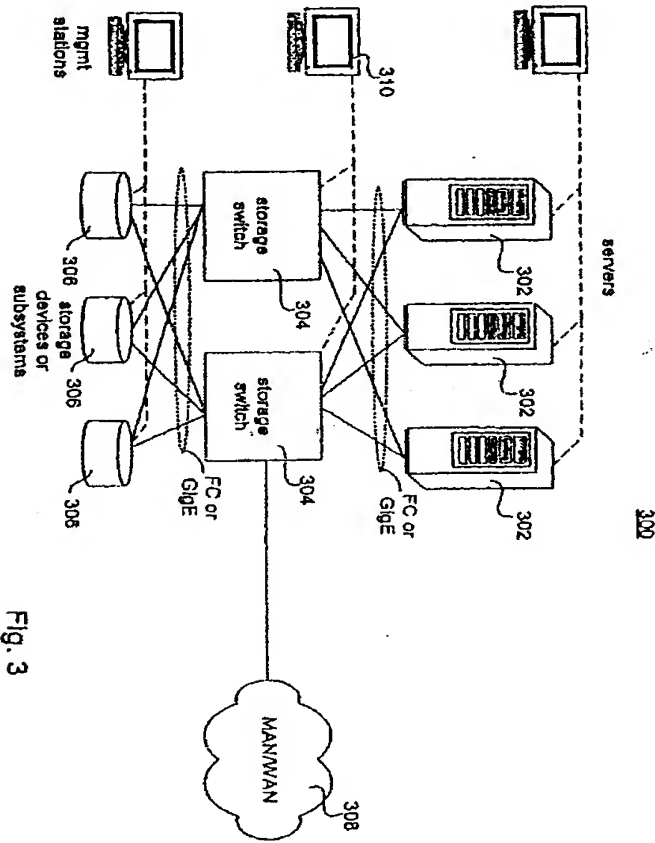
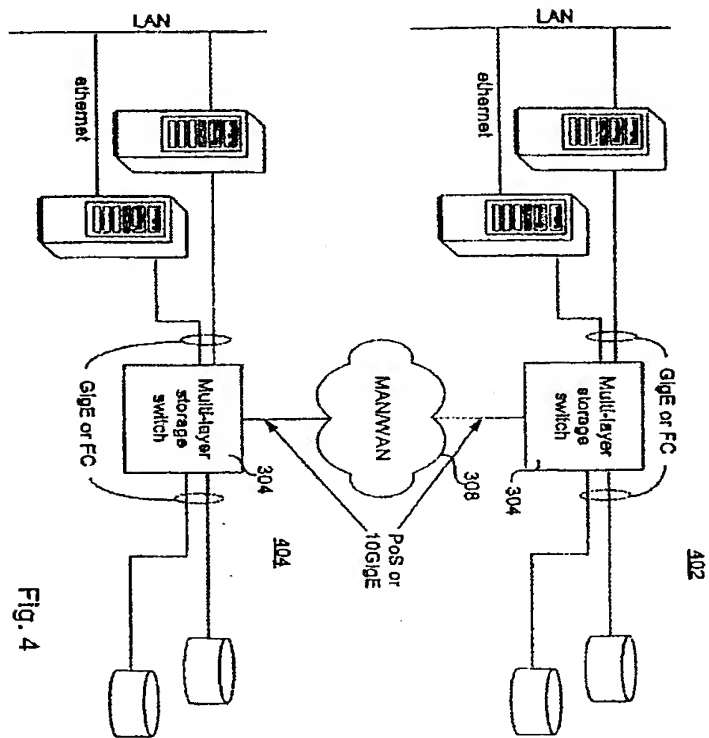


Fig. 3



**Fig. 4**

WO 03/027877

PCT/US02/30912

5/38

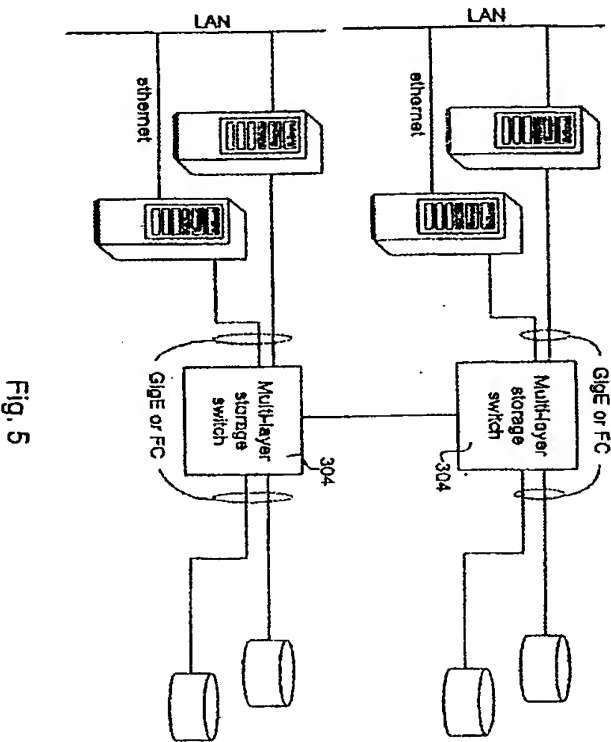


Fig. 5

WO 03/027877

PCT/US02/30912

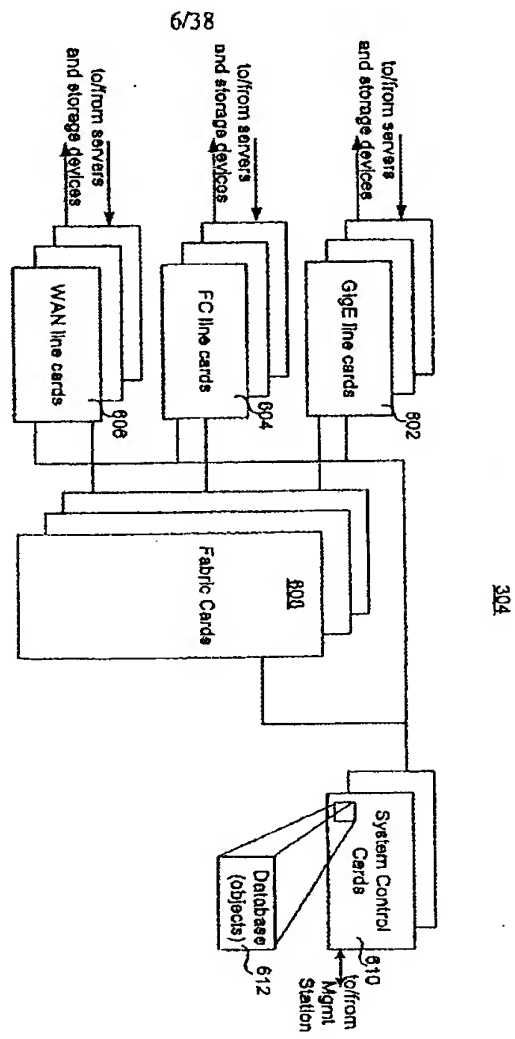


Fig. 6

WO 03/027877

PCT/US02/30912

7/38

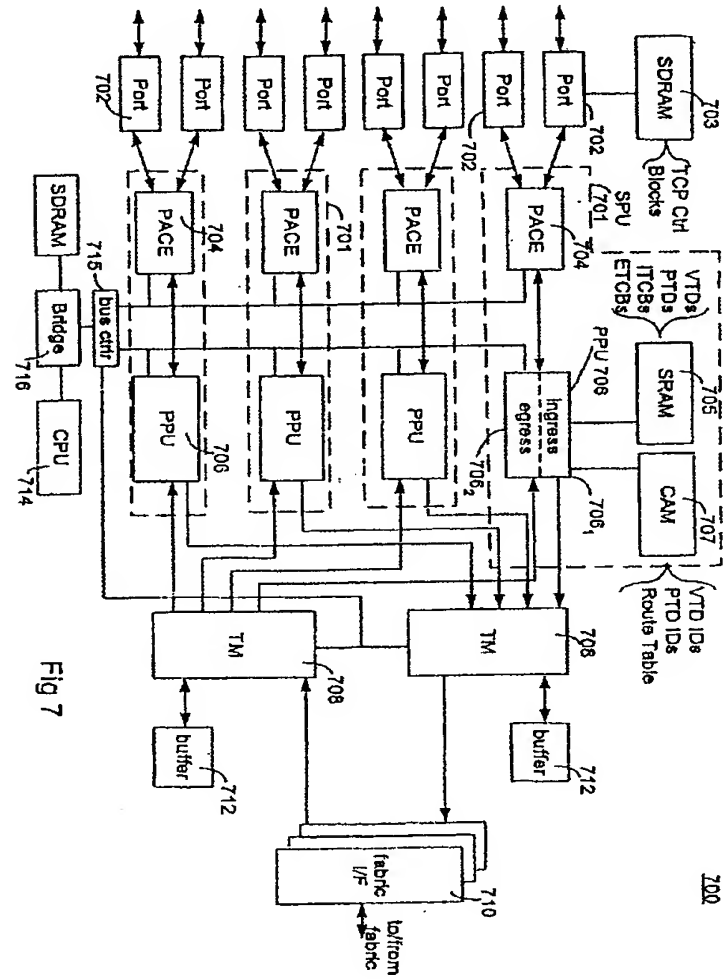


Fig 7

WO 03/027877

PCT/US02/30912

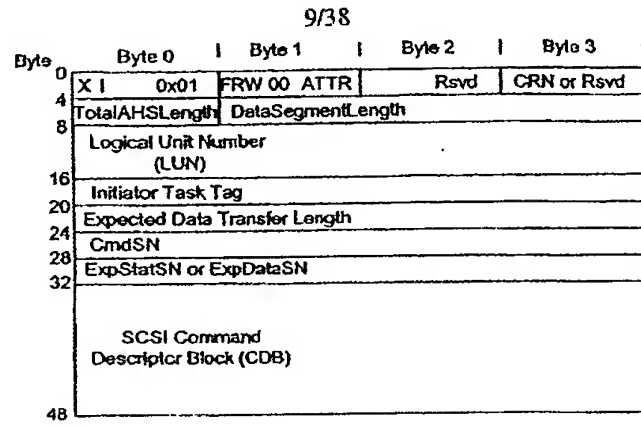
8/38

VTD	
VTD ID	
FlowID	
Extent Descriptors (e.g., size, location)	
# of outstanding commands	
Max # of commands	
Response time	
LUN	
TCP control block index	
S_ID/D_ID	
MaxCmdSN	
Total open sequences	

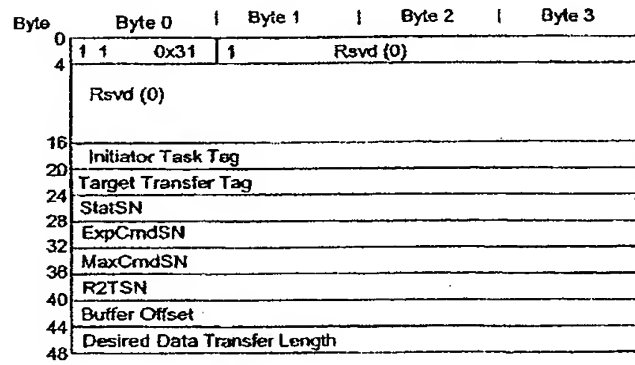
Fig. 7a

WQ 63/27877

PCT/US02/30912



ISCSI Command PDU  
Fig. 8a



ISCSI R2T PDU  
Fig. 8b

WO 03/027877

PCT/US02/30912

10/38

Byte	Byte 0	Byte 1	Byte 2	Byte 3
0	0 0	0x05	F	Rsvd (0)
4	Rsvd (0)	DataSegmentLength		
8	LUN or Reserved (0)			
16	Initiator Task Tag			
20	Target Transfer Tag or 0xffffffff			
24	Rsvd (0)			
28	ExpCmdSN			
32	Rsvd (0)			
36	DataSN			
40	Buffer Offset			
44	Rsvd (0)			
48	Data			

ISCSI Write Data PDU  
Fig. 8c

Byte	Byte 0	Byte 1	Byte 2	Byte 3
0	1 1	0x25	F O U S	Rsvd (0) Status or Rsvd
4	Rsvd (0)	DataSegmentLength		
8	Rsvd (0)			
16	Initiator Task Tag			
20	Rsvd (0)			
24	StatSN or Rsvd (0)			
28	ExpCmdSN			
32	MaxCmdSN			
36	DataSN			
40	Buffer Offset			
44	Residual Count			
48	Data			

ISCSI Read Data PDU  
Fig. 8d



WO 03/027877

PCT/US02/30912

11/38

	Byte 0	Byte 1	Byte 2	Byte 3
0	1 1 0x21	1 rsv 0 u 0 u 0	Status	Response
4	Rsvd (0)	DataSegmentLength		
8	Rsvd (0)			
16	Initiator Task Tag			
20	Basic Residual count			
24	StatSN			
28	ExpCmdSN			
32	MaxCmdSN			
36	ExpDataSN or Rsvd (0)			
40	ExpR2TSN or Rsvd (0)			
44	Bidi-Read Residual Count			
48	Sense Data and Response Data (optional)			

iSCSI Response PDU

Fig. 8e

WO 03027877

FCT/US02/30912

12/38

Bits Word	31-24	23-16	15-08	07-00
0	R_CTL	D_ID		
1	rsvd	S_ID		
2	TYPE	F_CTL		
3	SEQ_ID	DF_CTL	SEQ_CNT	
4	OX_ID		RX_ID	
5	RLTV_OFF			

FC Frame Header  
Fig. 8f

Field Name	Description	Size
FCP_LUN	logical unit number	8 bytes
FCP_CNTL	control field	4 bytes
FCP_CDB	SCSI command descriptor block	16 bytes
FCP_DL	Data Length	4 bytes

FCP\_CMND Payload  
Fig. 8g

WO 03/027677

FCT/US02/00912

13/38

Field Name	Description	Size
DATA_RO	Relative offset of first byte of FCP_DATA IU that follows	4 bytes
BURST_LEN	length of FCP_DATA IU that follows	4 bytes
rsvd		4 bytes

FCP\_XFR\_RDY Payload

Fig. 8h

Field Name	Description	Size
rsvd		4 bytes
rsvd		4 bytes
FCP_STATUS	field validity and SCSI status	4 bytes
FCP_RESID	residual count	4 bytes
FCP_SNS_LEN	Length of FCP_SNS_INFO field	4 bytes
FCP_RSP_LEN	Length of FCP_RSP_INFO field	4 bytes
FCP_RSP_INFO	FCP response info	m bytes
FCP_SNS_INFO	FCP sense info	n bytes

FCP\_RSP Payload

Fig. 8i

WO 03/027877

PCT/US02/30912

14/38

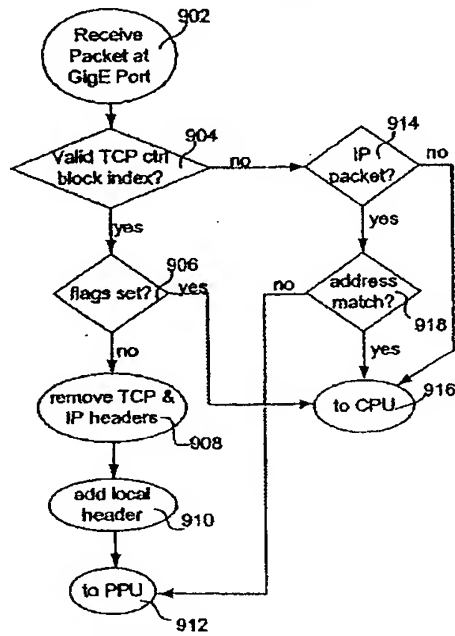
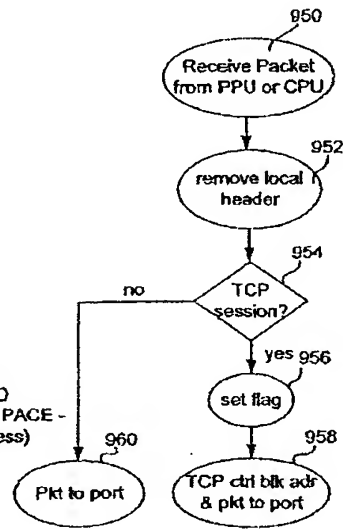


Fig 9a  
(Classification - PACE - iSCSI - Ingress)

Fig 9b  
(Classification - PACE - iSCSI - egress)



WO 03/027877

PCT/US02/30912

15/38

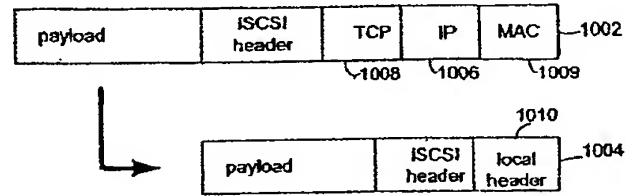


Fig. 10a

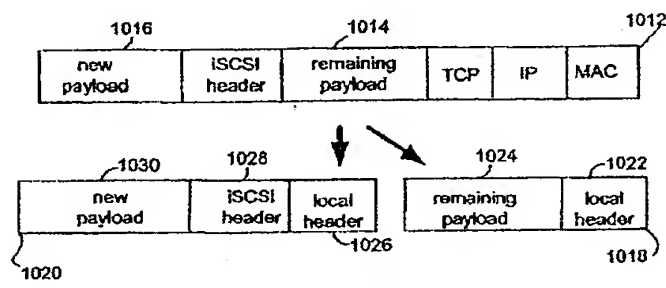


Fig. 10b

WO 03/027877

PCT/US02/06912

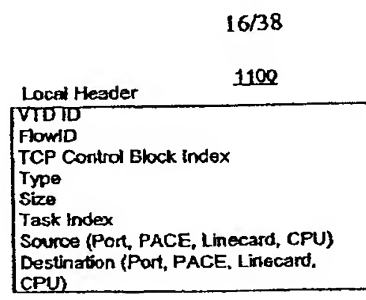


Fig. 11

WO 03/027877

PCT/US02/30912

17/38

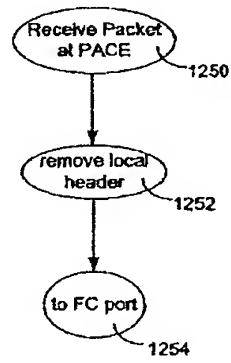
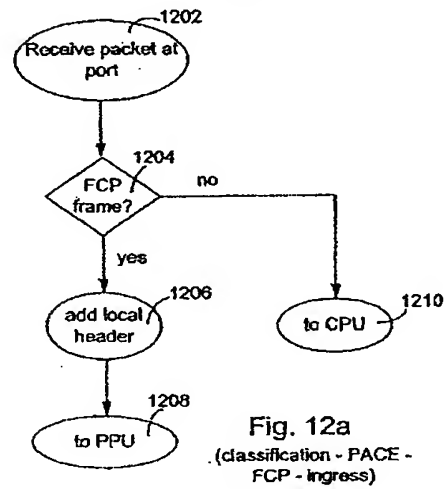


Fig. 12b  
(classification - PACE - FCP - egress)

WO 03/027977

PCT/US02/30912

18/38

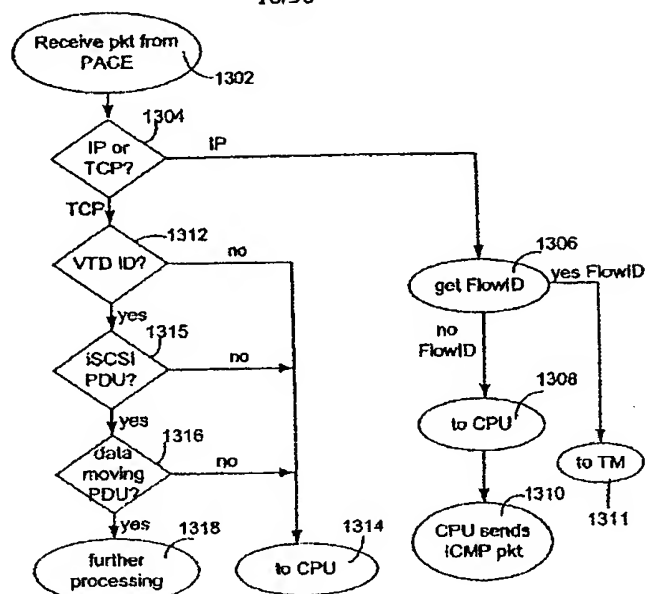


Fig. 13a  
(Classification - PPU -  
ingress)

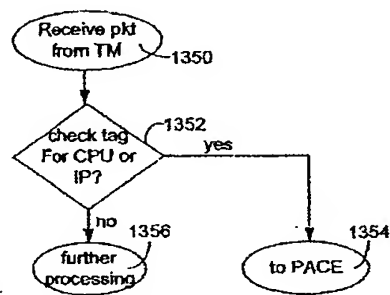


Fig. 13b  
(Classification - PPU -  
egress)



WO 03/027877

PCT/US02/00912

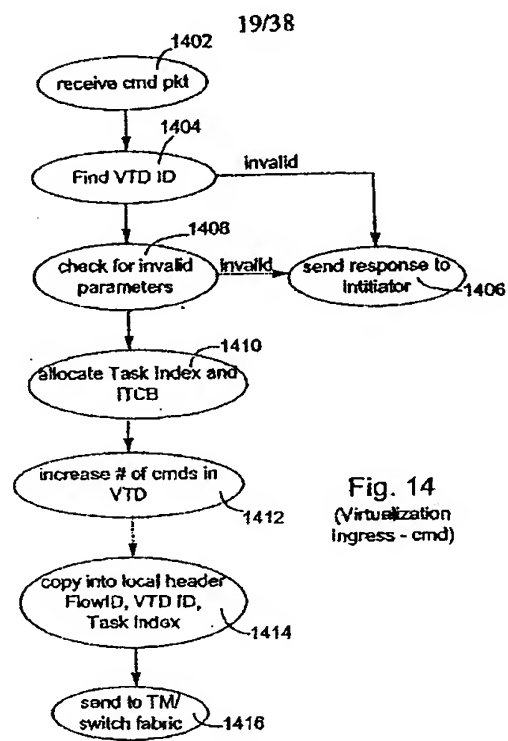


Fig. 14  
(Virtualization  
Ingress - cmd)

WO 03/027877

PCT/US02/30912

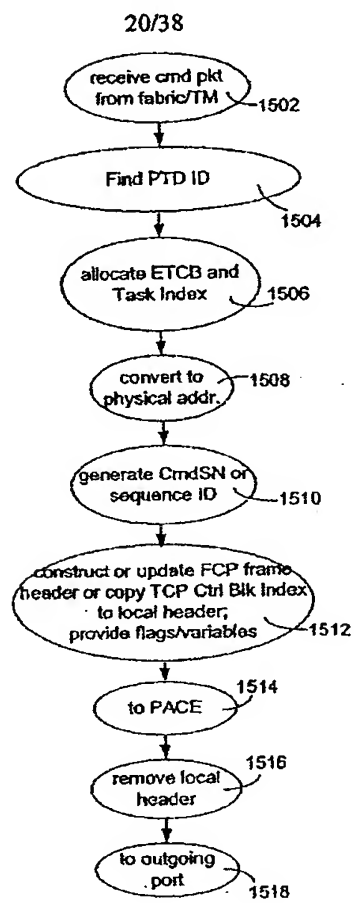


Fig. 15  
(Virtualization -  
Egress - cmd)

WO 03/027877

PCT/US02/30912

21/38

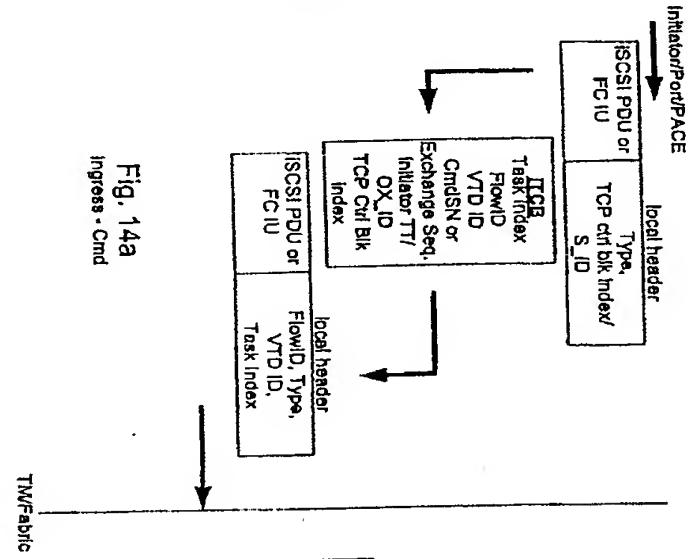


Fig. 14a  
Ingress - Cmd

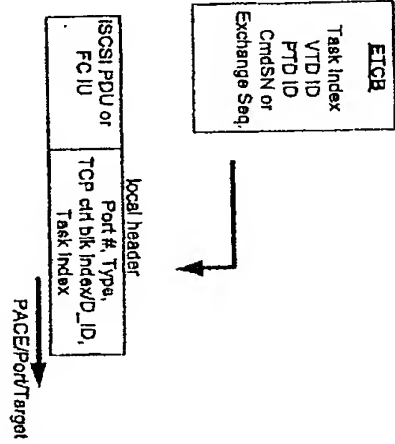


Fig. 15a  
Egress - Cmd

WO 03/027877

PCT/US02/30912

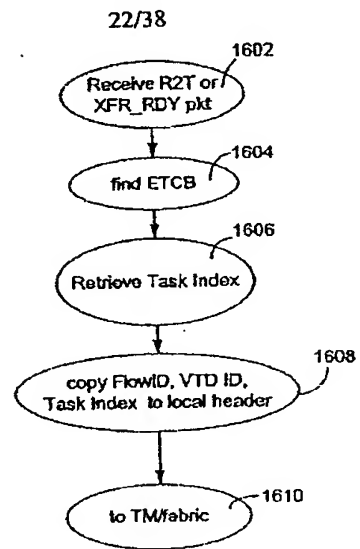


Fig. 16  
(Virtualization - Ingress -  
R2T/XFR\_RDY)

WO 03/027877

PCT/US02/30912

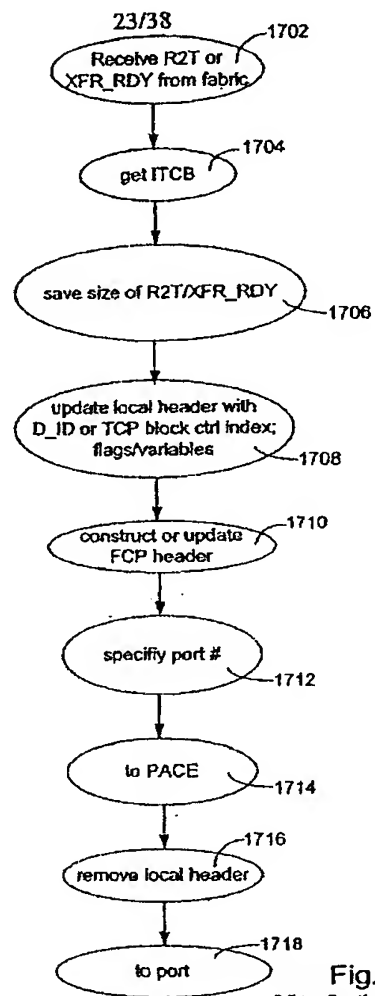


Fig. 17  
(Virtualization - Egress -  
R2T/XFR\_RDY)

WO 03/027877

PCT/US02/30912

24/38

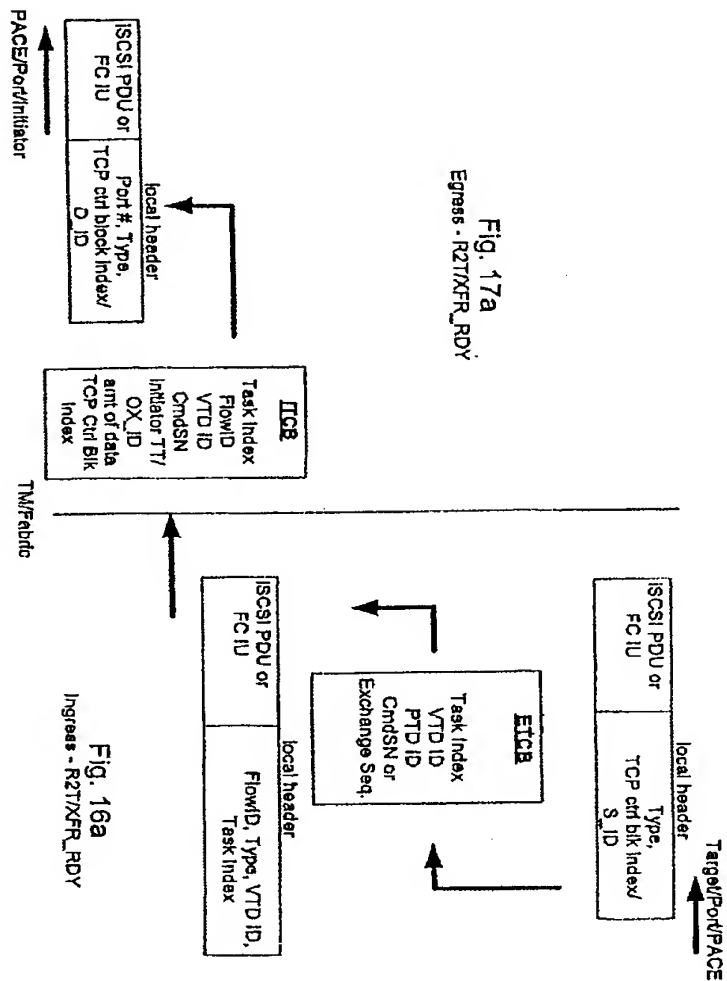


Fig. 16a  
Ingress - R2TXFR\_RDY

WO 03/027877

PCT/US02/38912

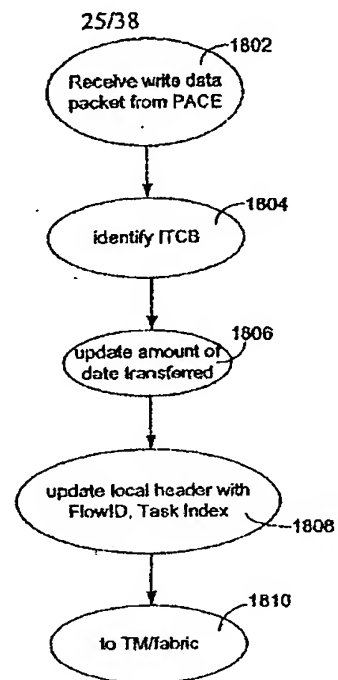


Fig. 18  
(Virtualization - Ingress -  
write data packet)

WO 03/027877

PCT/US02/30912

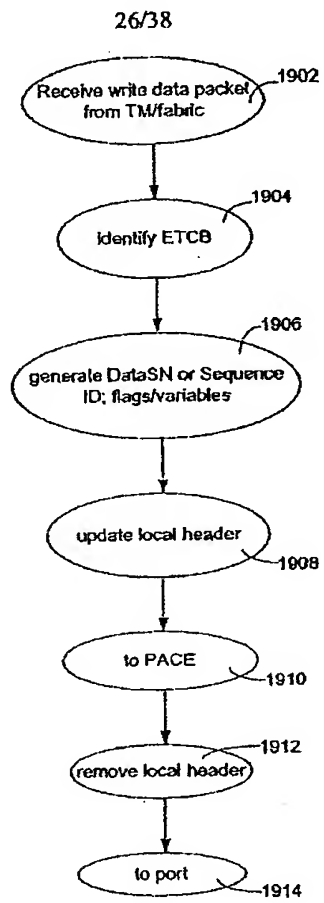


Fig. 19  
(Virtualization - Egress -  
write data pkt)



WO 63/627877

PCT/US02/30912

27/38

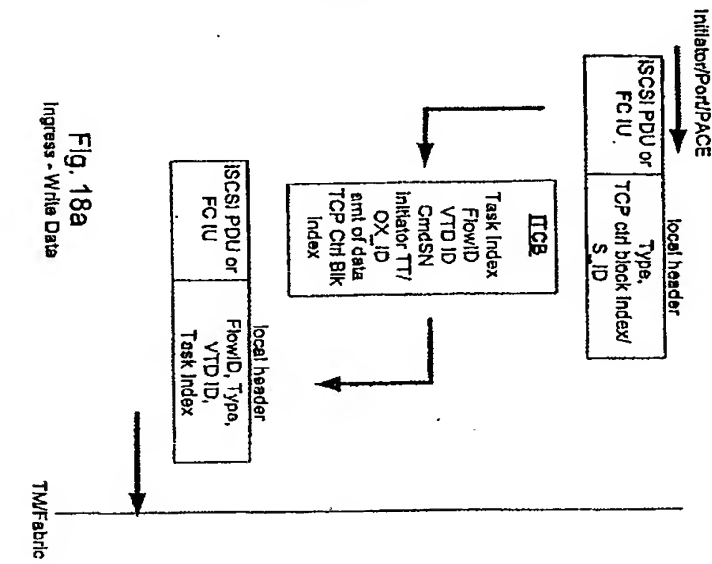


Fig. 18a  
Ingress - Write Data

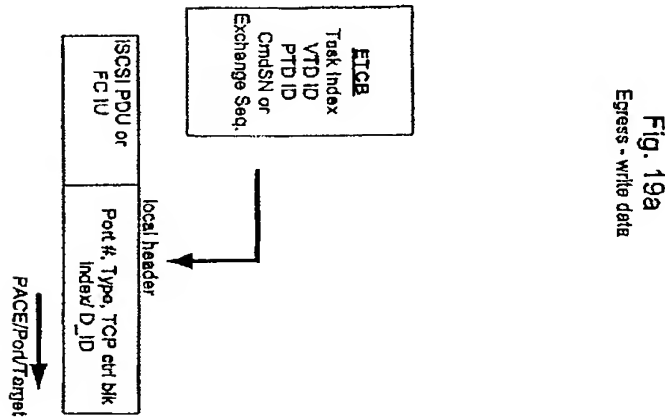


Fig. 19a  
Egress - write data

WO 03/027877

PCT/US02/30912

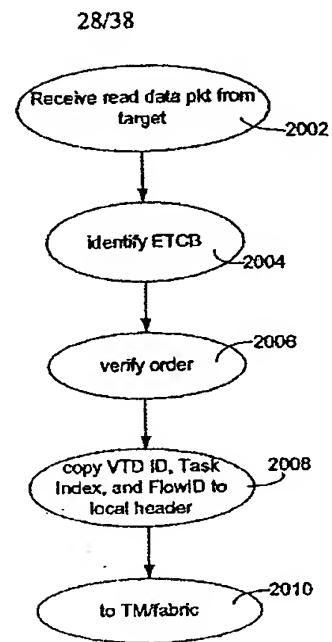


Fig. 20  
(Virtualization - Ingress -  
Read Data pkt)

WO 02/27877

PCT/US02/30912

29/38

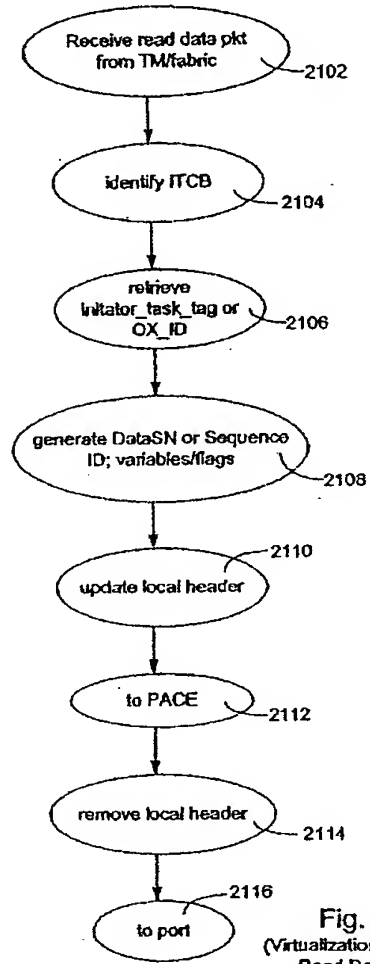


Fig. 21  
(Virtualization - Egress-  
Read Data pkt)

PCT/US02/30912

WO 03/027877

30/38

Fig. 21a  
Egress - Read Data

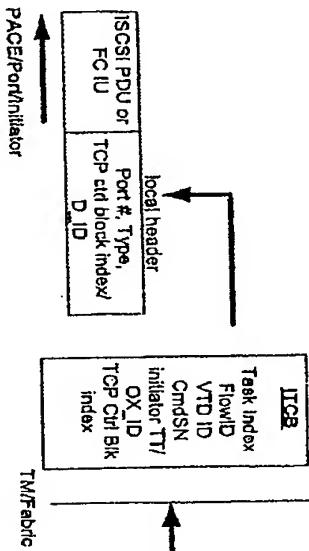
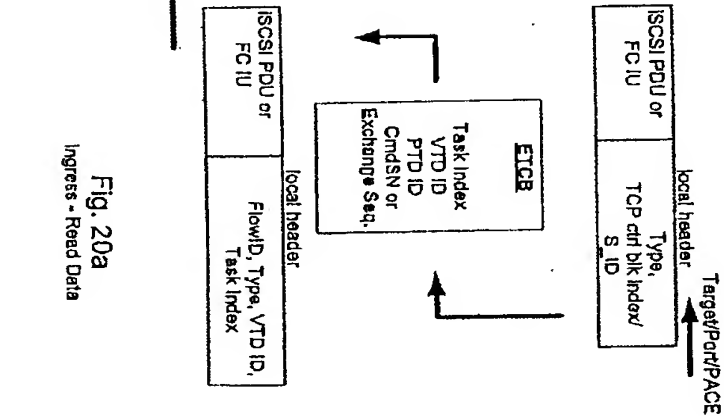


Fig. 20a  
Ingress - Read Data



WO 03/027677

PCT/US02/30912

31/38

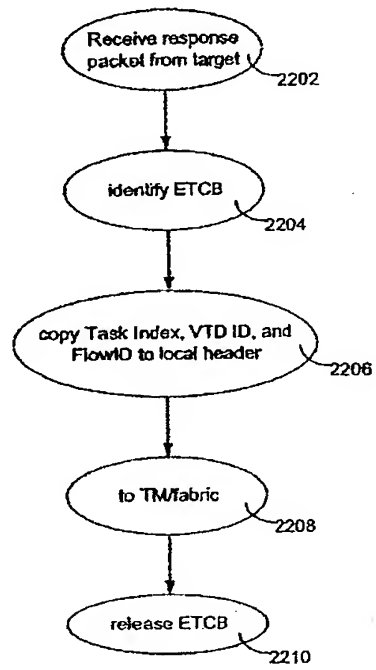
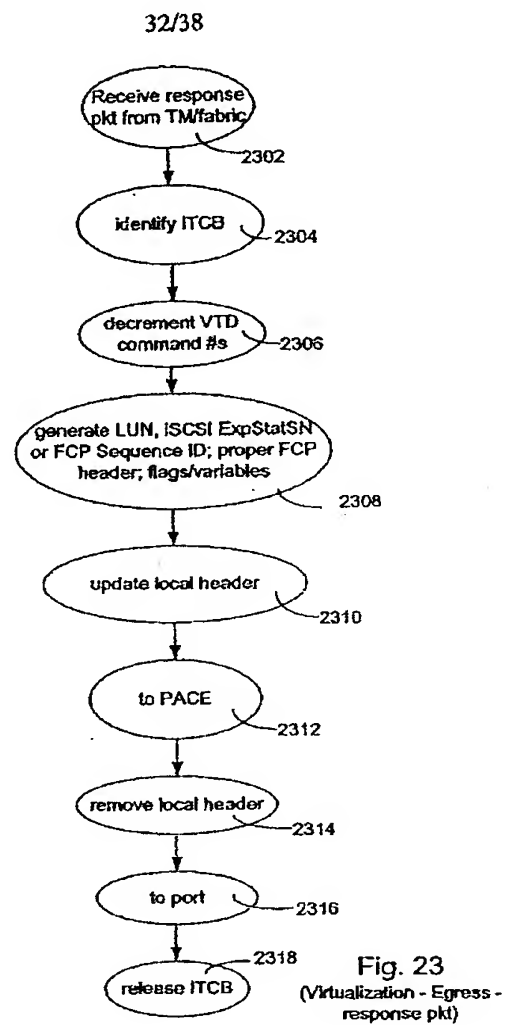


Fig. 22  
(Virtualization - Ingress -  
response pkt)

WO 03/027877

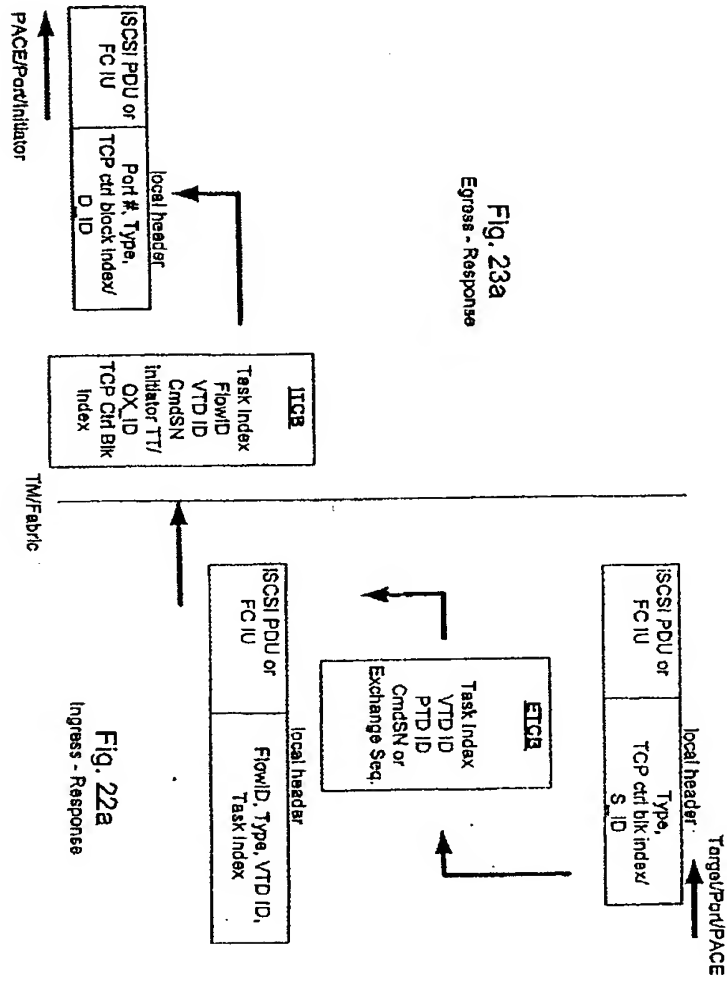
PCT/US02/30912



WO 03/027877

PCT/US02/30912

33/38



**Fig. 22a**  
Ingress - Response

WO 03/027877

PCT/US02/30912

34/38

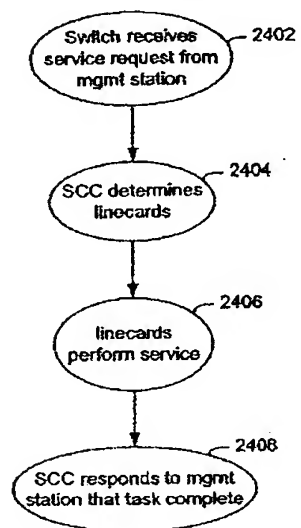


Fig.24



WO 03/027877

PCT/US02/30912

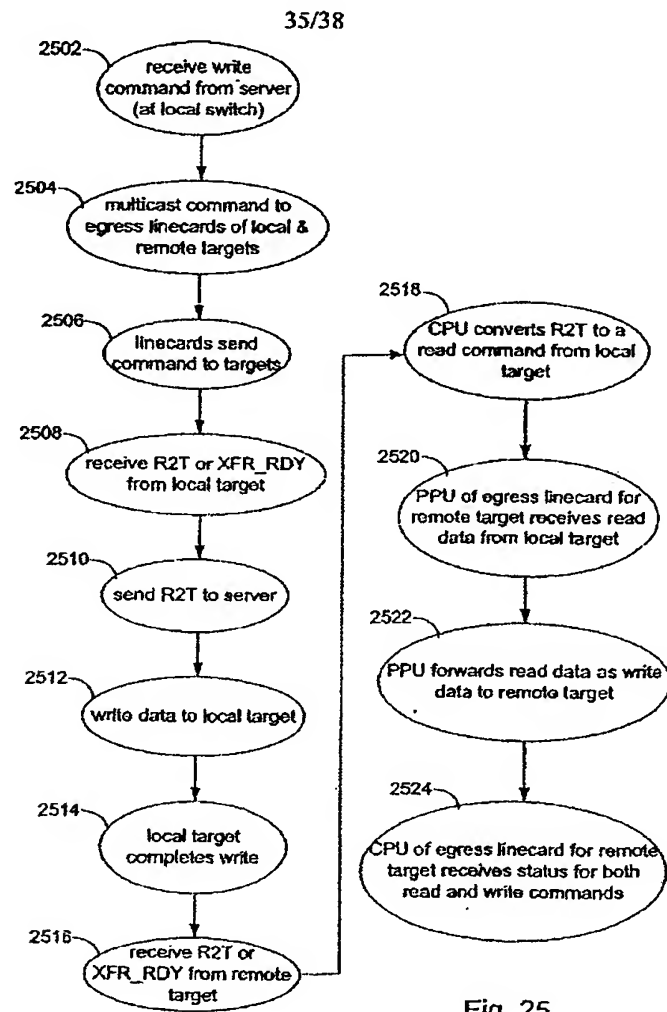


Fig. 25

WO 03/027877

PCT/US02/30912

36/38

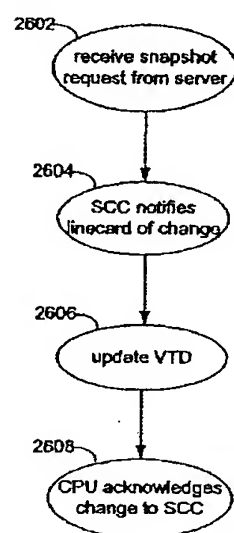


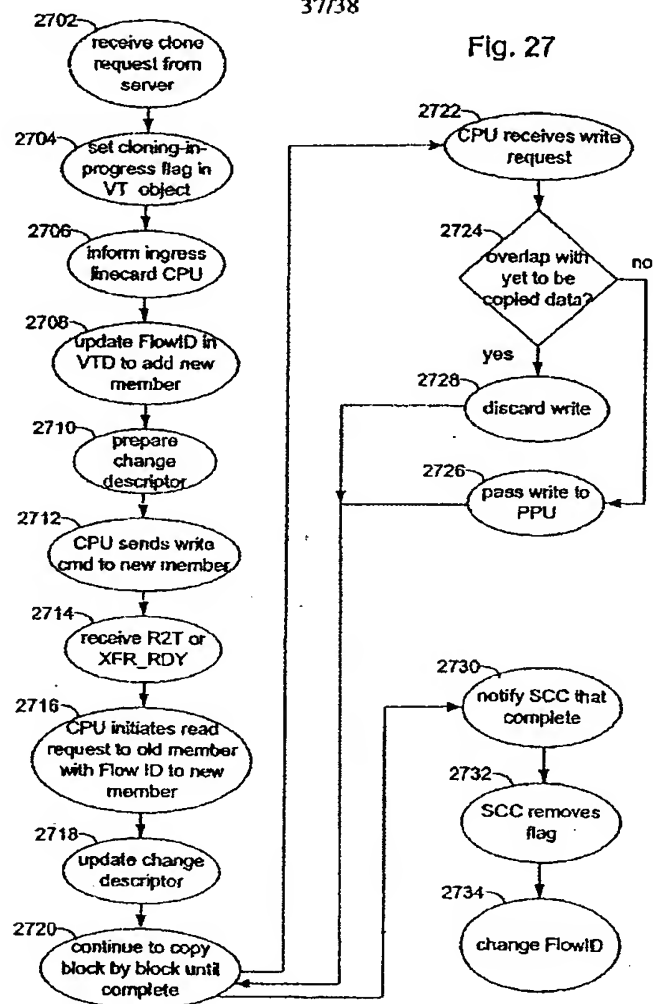
Fig. 26

WO 03/027877

PCT/US02/30912

37/38

Fig. 27



WO 03/027877

PCT/US02/00912

38/38

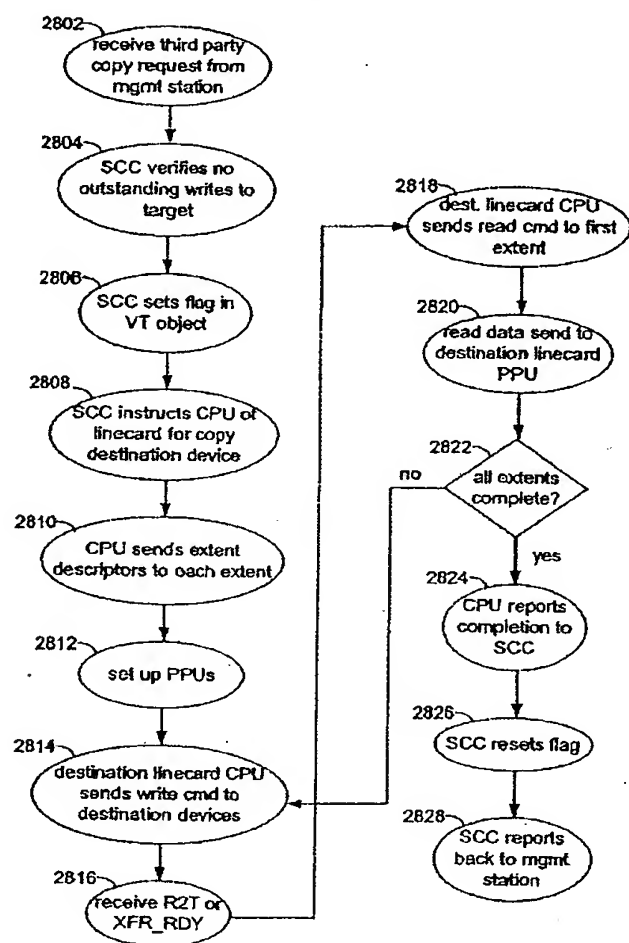


Fig. 28

## 【国際調査報告】

INTERNATIONAL SEARCH REPORT		International application No. PCT/US02/00912
<b>A. CLASSIFICATION OF SUBJECT MATTER</b> IPC(7) : G06F 15/16, G06F 9/00, G06F 15/177, H04L 12/56, H04L 3/16 US CL : 709/230, 230, 106, 200, 100, 101, 226-229, 230; 370/466, 465, 395, 402, 409, 355, 401 According to International Patent Classification (IPC) or to both national classification and IPC		
<b>B. FIELD(S) SEARCHED</b> Minimum documentation searched (classification system followed by classification symbols) U.S. : 709/230, 220, 106, 200, 100, 101, 226-229, 230; 370/466, 465, 395, 402, 409, 355, 401		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched EAST: USPAT, US-PGPH, EPO, JPO, IBM-TDB		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) fibre channel, iSCSI packet, format, protocol, translate, rep		
<b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y, P	U.S. 6,400,730 B1 (Latif et al.) 04 June 2002. Abstract col. 1, lines 19-35, col. 2, line 15 - col. 4, line 49 col. 6, line 44 - col. 8, line 9	1-49
A	U.S. 5,436,886 A (McGill) 25 July 1993 col. 3, line 57 - col. 5, line 12 col. 6, line 51 - col. 7, line 4	34, 37, 43, 48
Y, P	U.S. 2002/009797 A1 (Merrell et al.) 25 July 2002 Abstract pages 1-2; paragraphs 6, 10, pages 3-4, paragraphs 38	1-33, 35-36, 38-42, 44-47
Y, P	U.S. 6,434,656 B1 (Downer et al.) 13 August 2002 col. 4, lines 15-49, col. 6, line 45 - col. 7, line 34 U.S. 6,493,750 B1 (Mathew et al.) 10 December 2002 Abstract, col. 1, lines 19-34, col. 2, lines 5-43, col. 4, lines 10-61, col. 5, line 63 - col. 8, line 36	1-33 1-49
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.		
Special categories of cited documents: "A" Document defining the general state of the art which is not considered to be of particular relevance "E" Earlier application or patent published on or after the international filing date "L" Document which may appear earlier on priority claim(s) or which is cited to establish the publication date of another claim(s) or other special reason (as specified) "D" Document referring to an oral disclosure, use, exhibition or other means "P" Document published prior to the international filing date but later than the priority date claimed "X" Document of non-relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" Document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "A" Document member of the same patent family		
Date of the actual completion of the international search 23 January 2003 (23.01.2003)		Date of mailing of the international search report 28 FEB 2003
Name and mailing address of the ISA/US Communications and Trademarks Box PCT Washington, D.C. 20531 Facsimile No. (703)305-3230		Authorized officer Mark Powell James R. Mettling Telephone No. (703)305-9000

Form PCT/ISA/210 (second sheet) (July 1998)

## フロントページの続き

(81) 指定国 AP (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), EA (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), EP (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OA (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG), AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, N O, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW

(72) 発明者 ロレイアカー サントッシュ シー

アメリカ合衆国 カリフォルニア州 94086 サニーヴェイル #ジー 208 サウス フ  
ェア オークス アベニュー 655

(72) 発明者 チェン ユ ピン

アメリカ合衆国 カリフォルニア州 95120 サン ホセ ギャレット コート 1170

(72) 発明者 リー ワン フィ

アメリカ合衆国 カリフォルニア州 95035 ミルピタス オークハースト ウェイ 303

(72) 発明者 マラノン リネイト イー

アメリカ合衆国 カリフォルニア州 94550 リヴァーモア ノーマンディ サークル 22  
47

(72) 発明者 タン エンニュー

アメリカ合衆国 カリフォルニア州 95128 サン ホセ フルートデイル アベニュー 1  
919 アpartment イー213

Fターム(参考) 5B014 EB05

5B065 BA01 CA18 CA30 CE21 ZA08

5B082 HA05 HA07

5K034 AA02 AA10 AA20 BB01 BB06 CC01 DD03 EE11 FF02 GG03

HH12 HH25 HH42 HH61 MM24